НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
Інститут прикладного системного аналізу
Кафедра математичних методів системного аналізу

«До захисту допущено»
В. о. завідувача кафедри
_____ О.Л. Тимощук
«___»_____20__ р.

Дипломна робота

на здобуття ступеня бакалавра
з напряму підготовки 6.050101 «Комп'ютерні науки»
на тему: «Система підтримки прийняття рішень для прогнозування демографічних процесів»

Виконав:
студент IV курсу, групи КА-55
Гнатюк Марина Степанівна

_____
Керівник:
професор, д.т.н. Бідюк П.І.

_____
Консультант з економічного розділу:
доцент, к.е.н. Шевчук О.А. _____

Консультант з нормоконтролю:
доцент, к.т.н. Коваленко А. Є. _____

Рецензент:
проф., д. т. н. Теленик С.Ф. _____

Засвідчую, що у цій дипломній роботі
немає запозичень з праць інших авторів
без відповідних посилань.
Студент _____

Київ – 2019 року

РЕФЕРАТ

Дипломна робота: 109 с., 18 рис., 11 табл., 2 додатки, 19 джерел.

ДЕМОГРАФІЧНІ ПРОЦЕСИ, СИСТЕМА ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ, НЕЛІНІЙНІ НЕСТАЦІОНАРНІ ПРОЦЕСИ, ПРОГНОЗУВАННЯ

Дана робота присвячена дослідженню методів прогнозування демографічних процесів, аналізу переваг та недоліків вже існуючих інструментів прогнозування та розробці власної системи підтримки прийняття рішень.

Метою дипломної роботи є дослідження засад прогнозування демографічних процесів, а також створення системи підтримки прийняття рішень для отримання практичних результатів.

Об’єктом дослідження є демографічні процеси, статистичні дані щодо чисельності та тенденції розвитку популяції населення різних країн світу.

ABSTRACT

The work consist of 109 pages, 18 images, 14 tables, 2 appendices, 19 sources

The theme: Decision support system for forecasting demographic process.

DEMOGRAPHIC PROCESSES, DECISION SUPPORT SYSTEM, NONLINEAR NON-STATIONARY PROCESSES, FORECASTING

This work is devoted to the research methods of forecasting demographic methods, to analyze the advantages and disadvantages of existing programs and to develop own decision support system for forecasting demographic process.

The purpose of the thesis is to study the basic principles of forecasting demographic processes, as well as the development of decision support system for obtaining practical results.

The subject of the study are demographic processes, statistical data on the number and trends of population development in different countries of the world.

ЗМІСТ

ВСТУП

Статус країни на світовій арені часто визначається її економічними показниками, які у свою чергу тісно пов’язані із демографічними процесами. Демографія відіграє надзвичайно важливу роль як всередині держави, так і за її межами. Чисельність населення, тенденція його розвитку та якість стану наглядно вказує на можливі зміни напряму вектора ведення внутрішньої політики.

Для уникнення можливих економічних, соціальних чи інших проблем та підвищення рівню життя населення, важливим є можливість оцінки тенденції розвитку демографічного стану країни, що забезпечить проведення більш глибокого аналізу її розвитку, який в результаті допоможе у прийнятті правильних рішень.

Тому постає питання про забезпечення функціоналу, який би давав змогу отримати точний прогноз демографічних процесів для їх подальшого аналізу та використання. Сьогодні представлено достатньо багато програмних пакетів, які реалізовують прогнозування процесів різного роду. Проте, технології розвиваються і стає можливим пришвидшення часу обробки, мінімізація об’ємів використовуваних ресурсів та покращення точності результатів програмних продуктів.

Ще одним відкритим питанням залишається розробка максимально простого, зрозумілого інтерфейсу та гнучкої інтеграції з різними операційними системами, що спростить процес використання СППР.

РОЗДІЛ 1 ОСОБЛИВОСТІ РОЗВИТКУ ДЕМОГРАФІЧНИХ ПРОЦЕСІВ ТА ПОБУДОВИ ЇХ МАТЕМАТИЧНИХ МОДЕЛЕЙ

1.1 Особливості розвитку демографічних процесів в Україні і Німеччині та задачі стосовно їх аналізу

До найважливіших показників рівня розвитку будь-якої країни світу належать економічний та демографічний стани, їх ще називають якісними та кількісними характеристиками населення. Демографічні та економічні процеси взаємозв'язані, оскільки населення – це в своєму роді безцінне джерело ресурсів для праці, носій економічних відносин, який має великий вплив на формування та розвиток економіки країни. Трудовий потенціал населення, і, як наслідок, величина національного доходу, залежить від демографічних показників, тому надзвичайно важливим є вміння якомога точніше зробити прогноз їх змін, щоб мати змогу вчасно прийняти правильне рішення.

За роки свого незалежного існування Україна завжди орієнтувалась на європейський рівень показників у економічній, соціальній, культурній, політичній та інших, не менш важливих, сферах життєдіяльності суспільства. Якщо ж порівнювати демографічний стан України та однієї з найпотужніших країн Європейського Союзу – Німеччини, то за останні двадцять років ситуація значно змінилась.

Проілюструємо показники двох обраних держав у розрізі наступних критеріїв: народжуваність, смертність, середня тривалість життя, чисельність міжнародної міграції, ВВП на душу населення(дол. США за поточним курсом). Рівень народжуваності – показник, що визначає середню кількість народжених на одну жінку. Від початку XXI ст. тенденція народжуваності як в Україні, так і в Німеччині майже не змінюється, як показано на рисунку 1.1.

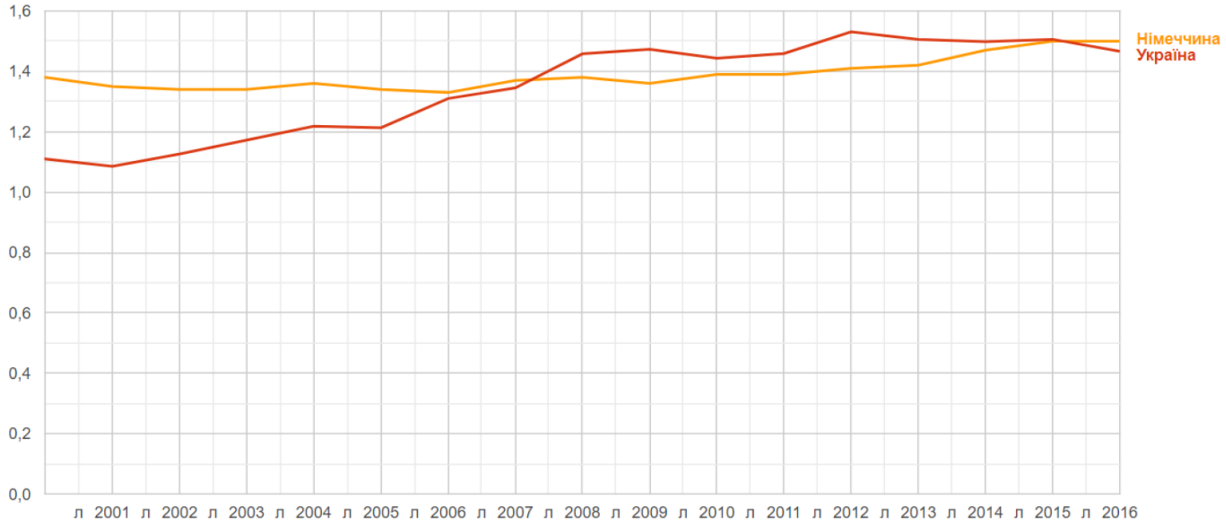


Рисунок 1.1 - Рівень народжуваності в Україні та Німеччині

Можна помітити, що рівень народжуваності в Україні змінився з 1.11 до 1.47, а в Німеччині з 1.38 до 1.5, що вказано у таблиці 1.1.

На даний момент все далі і далі відходять традиції багатодітності у нашій країні.

Сумарний показник народжуваності відносно минулого століття дедалі зменшився, що означає неможливість розширеного відтворення населення. З падінням рівню життя, що набуває все більшого розмаху, нікуди не зниклими наслідками Чорнобильської катастрофи та загостренням соціально-психологічних проблем, питання про розширення сім'ї відходить на інший план.

Зниження народжуваності – нормальна тенденція для багатьох розвинутих країн світу, таких як Німеччина, проте не можна забувати про ту межу, за якою стоїть процес депопуляції, який може виявитися незворотним.

Ще один фактор впливу зменшення народжуваності – це зменшення кількості шлюбів та збільшення кількості розлучених, бездітних, удівців, що ще більше погіршує демографічний стан.

Таблиця 1.1 - Рівень народжуваності в Україні та Німеччині.

рік	Україна	Німеччина
2000	1.11	1.38
2001	1.09	1.35
2002	1.13	1.34
2003	1.17	1.34
2004	1.22	1.36
2005	1.21	1.34
2006	1.31	1.33
2007	1.35	1.37
2008	1.46	1.38
2009	1.47	1.36
2010	1.44	1.39
2011	1.46	1.39
2012	1.53	1.41
2013	1.51	1.42
2014	1.5	1.47
2015	1.51	1.5
2016	1.47	1.5

Рівень смертності – показник, що визначає імовірність смерті на 1000 осіб. В Україні станом на 2013 рік рівень смертності серед чоловіків становить 292.15, серед жінок – 111.32 як показано на рисунках 1.2-1.3, що значно більше, ніж в Німеччині, дані вказані у таблиці 1.2.

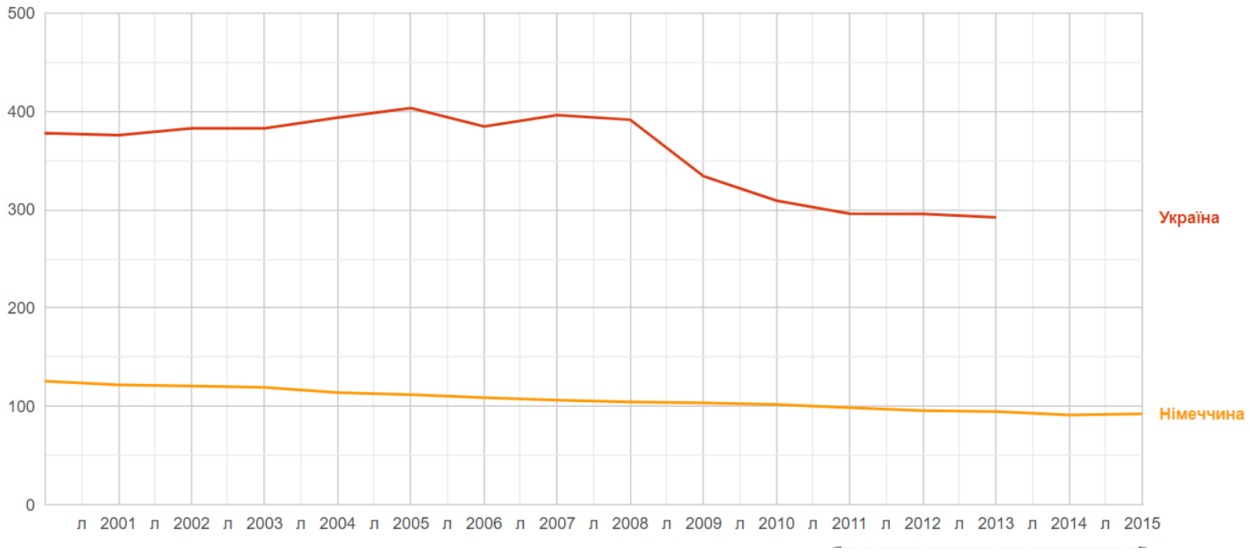


Рисунок 1.2 - Рівень смертності серед чоловіків в Україні та Німеччині

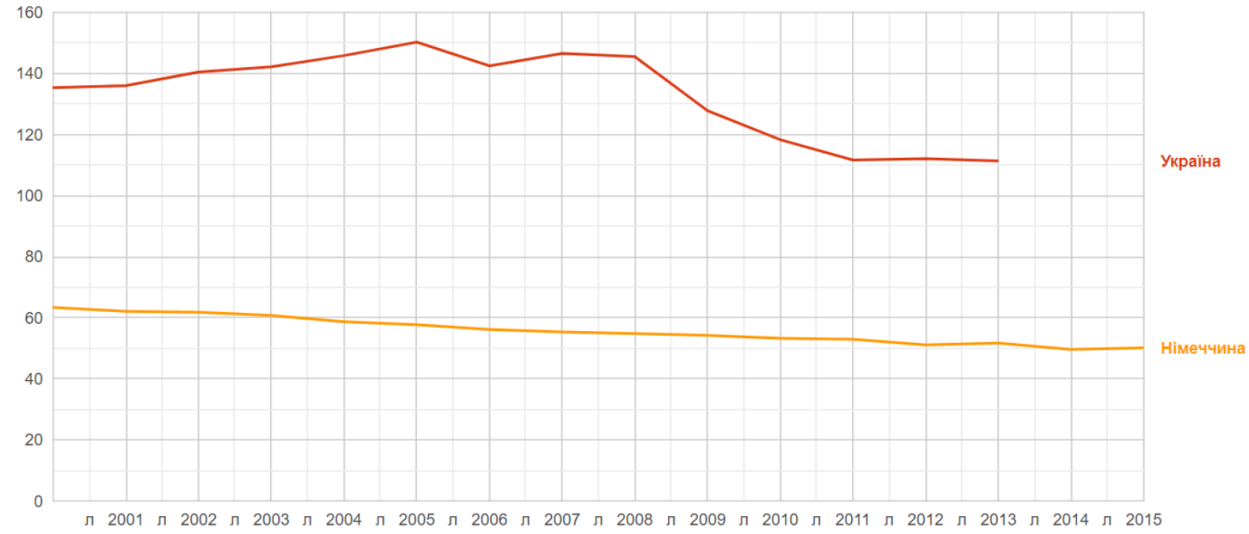


Рисунок 1.3 - Рівень смертності серед жінок в Україні та Німеччині

Не можна не помітити факт постійного росту рівня смертності серед українців працездатного віку. Так смертність чоловіків більша у 2,6 рази за смертність жінок. Німеччина ж має набагато менші показники смертності, ніж наша країна.

До одних з основних причин цього процесу належить втрата здоров’я через несприятливі умови життя і праці більшої частини населення, а також поширення шкідливих звичок, неефективна система охорони здоров’я та недотримання норм здорового способу життя.

Таблиця 1.2 - Рівень смертності в Україні та Німеччині

	Україна		Німеччина	
рік	чоловіки	жінки	чоловіки	жінки
2000	377.81	135.35	125.48	63.37
2001	375.82	136.03	121.89	62.09
2002	382.79	140.49	120.79	61.77
2003	382.68	142.19	119.39	60.75
2004	393.62	145.85	114.01	58.67
2005	403.5	150.23	111.85	57.72
2006	384.63	142.46	108.87	56.19
2007	396.21	146.58	106.21	55.3
2008	391.51	145.51	104.42	54.79
2009	334.1	127.78	103.62	54.25
2010	309.24	118.28	101.93	53.27
2011	296.05	111.64	98.66	52.93
2012	295.62	112.07	95.53	51.09
2013	292.15	111.32	94.66	51.72
2014			91.12	49.59
2015			92.46	50.17

Середня тривалість життя – показник що характеризує тривалість життя людей в середньому. Середній вік чоловіків в Україні становить 66 років, жінок – 76 років, що зображено на рисунках 1.4-1.5. В Німеччині цей показник сягає 78 і 83 відповідно, дані наведено в таблиці 1.3.

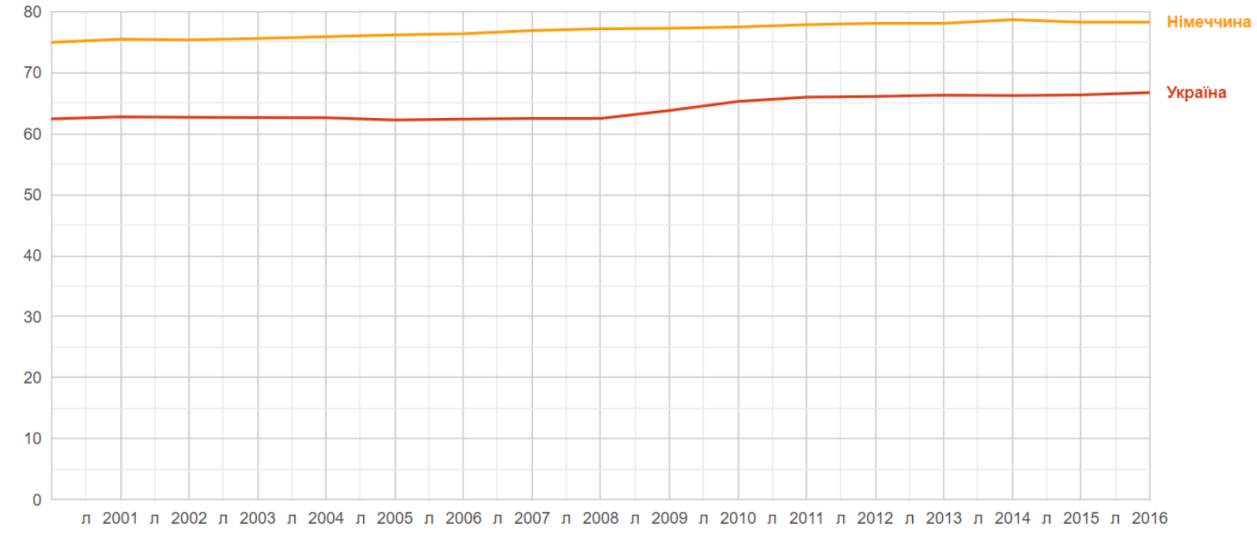


Рисунок 1.4 - Середня тривалість життя чоловіків в Україні та Німеччині

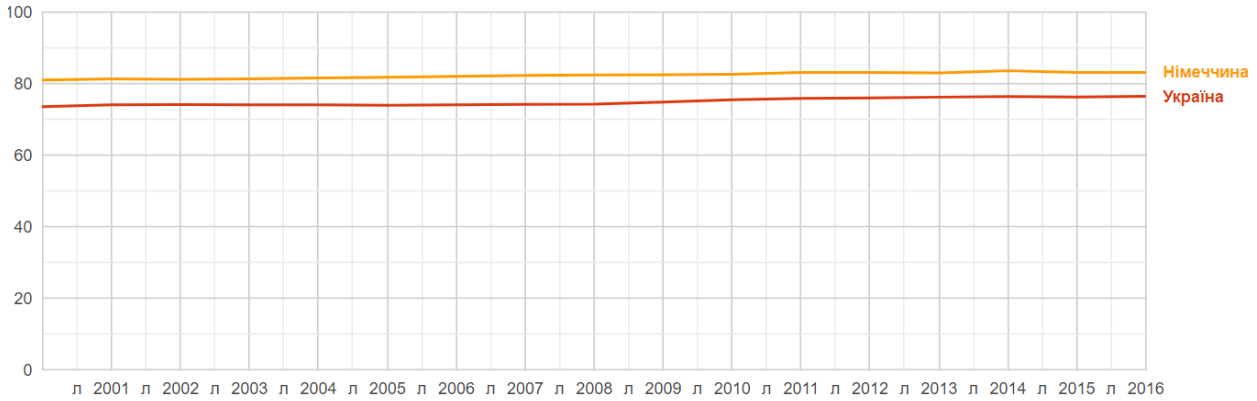


Рисунок 1.5 - Середня тривалість життя жінок в Україні та Німеччині

Таблиця 1.3 - Середня тривалість життя в Україні та Німеччині

	Україна		Німеччина	
рік	чоловіки	жінки	чоловіки	жінки

2000	62.44	73.55	75	81
2001	62.77	74.08	75.5	81.3
2002	62.7	74.13	75.4	81.2
2003	62.64	74.06	75.6	81.3
2004	62.6	74.05	75.9	81.6
2005	62.23	73.97	76.2	81.8
2006	62.38	74.06	76.4	82
2007	62.51	74.22	76.9	82.3
2008	62.51	74.28	77.2	82.4
2009	63.79	74.86	77.3	82.5
2010	65.28	75.5	77.5	82.6
2011	65.98	75.88	77.9	83.1
2012	66.11	76.02	78.1	83.1
2013	66.34	76.22	78.1	83
2014	66.25	76.37	78.7	83.6
2015	66.37	76.25	78.3	83.1
2016	66.73	76.46	78.3	83.1

Чисельність міжнародної міграції – це кількість людей, які проживають у країнах, уродженцями яких вони не є. Також до них належать біженці.

В Україні зафіксовано 4.8349 млн мігрантів, коли в Німеччині їх чисельність становить 12.0057, що майже втричі більше, що зображено на рисунку 1.6.

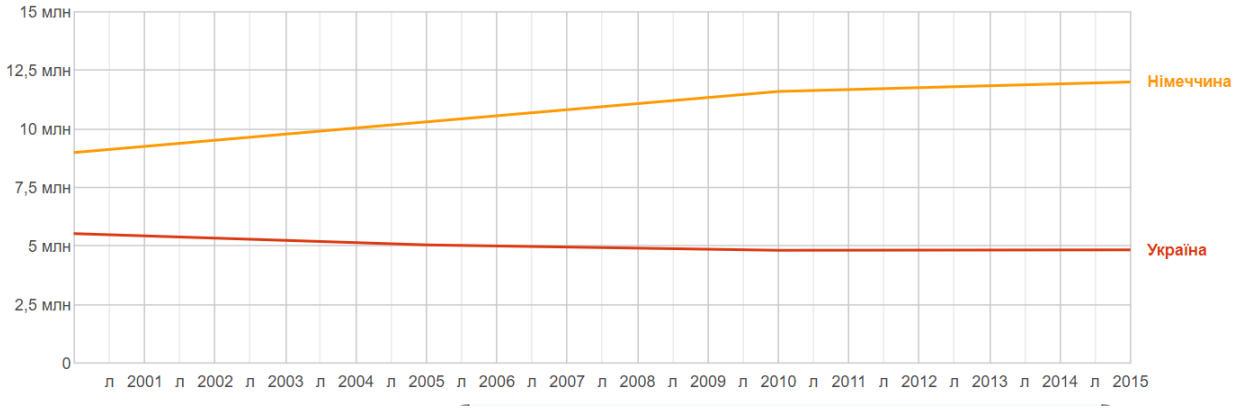


Рисунок 1.6 - Чисельність міжнародної міграції в Україні та Німеччині

Ще одним доленосним фактором демографічного стану є міграційні процеси. Зазвичай виділяють два основні види: переселення або ж проживання тривалого терміну та трудова міграція. Сьогодні особливою популярністю користується трудова міграція. Чи не у кожній другій сім'ї близькі вимушені покинути рідну країну в пошуках заробітку, який, на жаль, не можуть здобути тут. Слід звернути особливу увагу на показники, яких досягає Україна за кількістю своїх громадян, що виїжджають за її межі, і дуже часто не повертаються, оскільки умови життя та праці, рівень заробітку цілком переважає над тим, що можна отримати, залишившись. Також слід зазначити, що основна частина емігрантів – молоді люди, їх виїзд означає зменшення осіб працездатного віку, а отже погіршення вікової структури України.

Тенденція кількості людей, що вирішили іммігрувати до України з роками не збільшується, що зовсім протирічить показникам Німеччини, за останні роки кількість населення Німеччини зросла до рекордної кількості, завдяки іммігрантам, зокрема серед них чимало наших співвітчизників, дані наведено в таблиці 1.4.

Таблиця 1.4 - Чисельність міжнародної міграції в Україні та Німеччині

рік	Україна (млн)	Німеччина (млн)
2000	5.5271	8.9926
2005	5.0503	10.2992
2010	4.8188	11.6057
2015	4.8349	12.0057

ВВП на душу населення – це продукт, розділений на середньорічне населення. ВВП – це сума валової вартості, яку додають усі виробники-резиденти, плюс податки на продукцію, мінус будь-які субсидії, не включені до вартості продуктів. Показники ВВП України в порівнянні з Німеччиною надзвичайно малі, що видно на рисунку 1.7. Станом на 2017 рік вони різняться в 16,8 разів, дані наведено в таблиці 1.5.

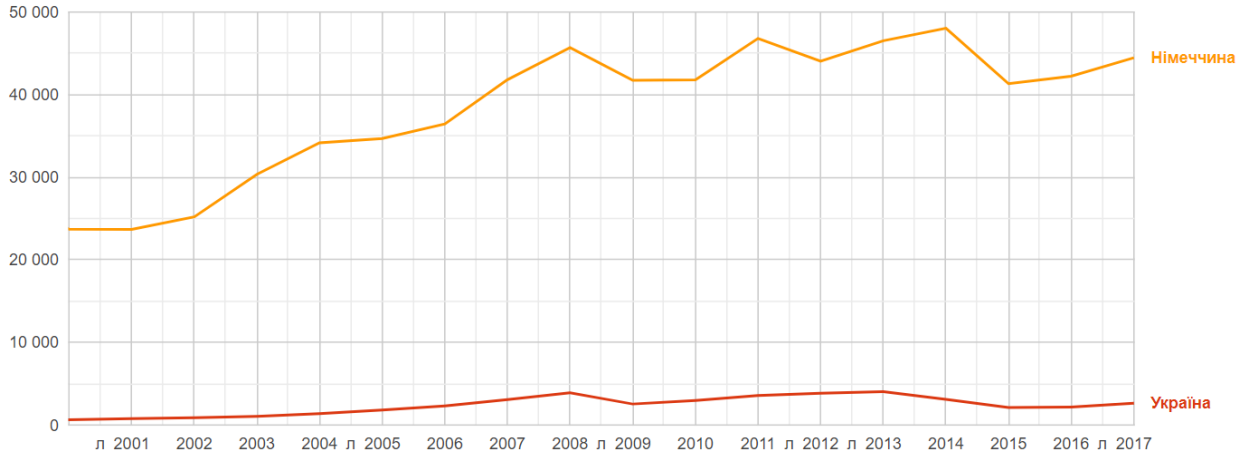


Рисунок 1.7 - ВВП на душу населення (дол. США за поточним курсом) в Україні та Німеччині

Таблиця 1.5 - ВВП на душу населення (дол. США за поточним курсом) в Україні та Німеччині

рік	Україна	Німеччина
2000	635.71	23718.75
2001	780.74	23687.32
2002	879.48	25205.16
2003	1048.52	30359.95
2004	1367.35	34165.93
2005	1828.72	34696.62
2006	2303.02	36447.87
2007	3068.61	41814.82
2008	3891.04	45699.2
2009	2545.48	41732.71
2010	2965.14	41785.56
2011	3569.76	46810.33
2012	3855.42	44065.25
2013	4029.72	46530.91
2014	3104.66	48042.56
2015	2124.66	41323.92
2016	2185.73	42232.57
2017	2639.82	44469.91

Також, ще одним надзвичайно важливим показником, є кількість смертей, пов’язаних з бойовими діями. Смерті, пов’язані з бойовими діями, означають смерті в пов’язаних із бойовими діями конфліктах між супротивними сторонами в конфліктній діаді (двома суб’єктами-сторонами конфлікту). Зазвичай смерті,

пов’язані з бойовими діями, трапляються на війні за участі збройних сил супротивних сторін. До таких бойових дій належать звичайні бої на полі бою, діяльність партизанських формувань і всі види бомбардування військових частин, міст, сіл тощо. Зазвичай бойові дії націлені на самих військових і їхні споруди або на державні установи й представників держави, проте часто спостерігаються значні супутні втрати через загибель цивільного населення під перехресним вогнем, під час нерозбірливого бомбардування тощо. Усі смерті (і військових, і цивільних осіб), що відбулися в таких ситуаціях, зараховуються як смерті, пов’язані з бойовими діями. І, на жаль, Україна має значні втрати за останні п’ять років, що зображено на рисунку 1.8.

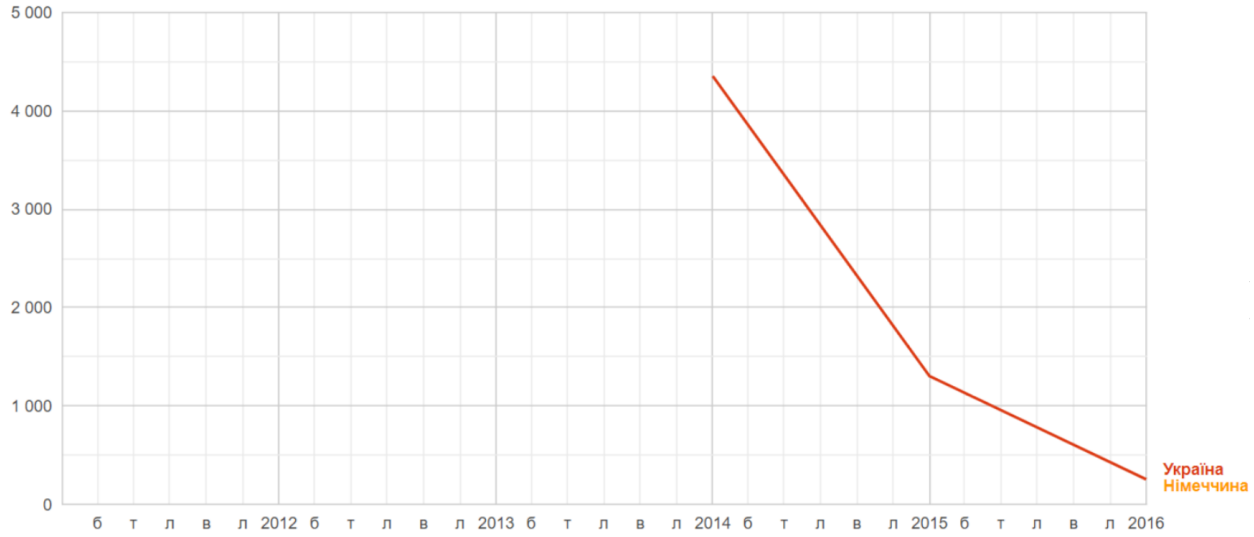


Рисунок 1.8 - Показники кількості смертей внаслідок бойових дій в Україні

Всі ці фактори мають неповоротній вплив на демографічний розвиток України та Німеччини. Демографічне питання поставало вже не раз, і тому існує різна варіація видів моделей в задачах аналізу, що перед собою ставили дослідники даного напряму. У роботах [4], [5], [6] можна знайти дані за додатковими критеріями.

Для дослідження демографічних процесів зазвичай використовують регресійний аналіз. Будують моделі парної регресії, нелінійної регресії, використовують нейромережі та байєсівські мережі. Серед існуючих, найбільш практично застосовуваними методами моделювання і прогнозування стаціонарних та нестаціонарних процесів, є методи експоненційного згладжування, АР, АРКС, моделі АРУГ і УАРУГ. Одним із конкретних прикладів є дослідження демографічних процесів з використанням моделі побудови прогнозу за допомогою використання оптимальних Байєсівських процедур розпізнавання на ланцюгах Маркова, що описано у роботах [5] та [13].

Тобто можемо зробити висновок, що дана проблема є актуальною для досліджень, і над її вдосконаленням працюють чимало людей, застосовуючи різноманітні методи моделювання.

1.2 Математичні моделі, що використовуються для опису демографічних процесів

При дослідженнях в області прогнозування зокрема демографічних процесів, користуються неабиякою популярністю регресійні, байєсівські моделі та нейромережі. Коротко розглянемо деякі з них, а більше детальну інформацію можна знайти у [1], [2], [3] роботах.

А. Авторегресія

Рівняння авторегресії описується впливом попередніх значень станів на поточний стан:

$$y(k) = a_0 + a_1 y(k-1) + \cdots + a_p y(k-p) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + \varepsilon(k),$$

де p – порядок авторегресії;

k - дискретний час;

$a_i -$ визначені на основі значень часового ряду, коефіцієнти моделі;

$\varepsilon(k) -$ випадкова величина, яка може з'являтись через похибки рівняння, вплив випадкових збурень, обчислювальні похибки тощо.

B. Парна регресія

До парної регресії входить так званий регресор, тобто незалежна змінна в правій стороні рівняння:

$$y(k) = a_0 + a_1 x(k) + \varepsilon(k),$$

де $x(k) -$ регресор

C. Множинна регресія

Множинна регресія дістала таку назву через те, що в її основі лежить відображення впливу на залежну змінну декількох незалежних:

$$y(k) = a_0 + a_1 x_1(k) + \cdots + a_p x_p(k) + \varepsilon(k)$$

де $x_1(k), \ldots, x_p(k) -$ регресори рівняння.

D. Змішана регресія

$$y(k) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + b_1 x_1(k) + \cdots + b_p x_p(k) + \varepsilon(k)$$

E. Авторегресія з ковзним середнім порядку (p,q)

$$y(k) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + \sum_{j=1}^{q} b_j \varepsilon(k-j) + \varepsilon(k)$$

F. Нелінійна регресія відносно змінних

$$y(k) = a_0 + a_1 x(k) + \cdots + a_p x^p(k) + \varepsilon(k)$$

Для оцінки коефіцієнтів псевдолінійної регресії, необхідно користуватись такими ж методами, тобто МНК та ММП. Запис даного рівняння нагадує поліноміальну регресію порядку p.

G. Нелінійна регресія відносно параметрів

$$y(k) = a_0 + a_1 e^{bx(k)} + \varepsilon(k)$$

У даній моделі, що нелінійна відносно параметрів, присутні добутки параметрів моделей, які включені до адитивних членів. Для оцінки цих параметрів слід користуватися МНК, ММП, Методом Монте-Карло, тощо.

H. Моделі гетероскедастичних процесів

Гетероскедастичними називаються моделі, дисперсія яких є змінна з часом. Рівняння для умовної дисперсії таких моделей має наступний вигляд:

$$h(k) = \beta_0 + \varepsilon^2(k-1) + \varepsilon_1(k),$$

де $\varepsilon^2(k) -$ квадрат залишків;

$h(k) -$ умовна дисперсія процесу;

$\varepsilon_1(k) -$ похибка моделі;

$k -$ дискретний час.

Узагальнена авторегресійна умовно гетероскедастична модель:

$$h(k) = \beta_0 + \sum_{i=1}^{p} \beta_i \varepsilon^2(k-i) + \sum_{i=1}^{q} a_i h(k-i) + \varepsilon_1(k),$$

де $\alpha, \beta > 0$ для уникнення від'ємних значень умовних дисперсій.

Авторегресійна умовно гетероскедастична модель порядку p

$$h(k) = \beta_0 + \sum_{i=1}^{p} \beta_i \varepsilon^2(k-i) + \varepsilon_1(k).$$

I. Байєсівська мережа

Байєсівська мережа описується у вигляді пари <G, B>, де G – модифікований нециклічний граф, а B – множина параметрів, за допомогою яких визначається мережа. Параметри цієї компоненти мають наступний вигляд:

$$\Theta_{X^{(i)}|pa(X^{(i)})} = P(X^{(i)}|pa(X^{(i)})),$$

де $x^{(i)} \in X^{(i)}$ та $pa(X^{(i)}) \in Pa(X^{(i)})$, $Pa(X^{(i)})$ – набір батьків змінної $X^{(i)}$.

Для отримання повної спільної ймовірності, необхідно скористатись формулою:

$$P_B\left(X^{(1)}, \dots, X^{(N)}\right) = \prod_{i=1}^{N} P_B(X^{(i)}|Pa(X^{(i)})).$$

Крім засобів знаходження умовних ймовірностей в орієнтованих графах існує чимало інших байєсівських методологій, таких як моделі динамічних процесів, моделі із симетричними зв'язками, моделі із прихованими змінними, що детальніше описані у роботі [13].

1.3 Статистичні тести для встановлення належності демографічних процесів до певного класу

1.3.1 Тести на стаціонарність

Крім визначення та побудови математичних моделей, ще одним важливим кроком є аналіз моделі, її корегування, визначення таких характеристик як наявність тренду та гетероскедастичності, для постановки наступних етапів прогнозування. Для перевірки процесу на стаціонарність рекомендується використовувати тест Дікі-Фуллера, що описаний у роботі [3].

Тест Дікі-Фуллера допомагає визначити, присутність одиничного кореню, тобто наявність нестаціонарності, шляхом використання трьох наступних рівнянь:

$$\Delta y(k) = \gamma y(k-1) + \varepsilon(k), \qquad (1.1)$$

$$\Delta y(k) = a_0 + \gamma y(k-1) + \varepsilon(k), \qquad (1.2)$$

$$\Delta y(k) = a_0 + \gamma y(k-1) + a_2 k + \varepsilon(k), \qquad (1.3.)$$

де k – дискретний час;

$\gamma = a_1 - 1$ – коефіцієнт в рівнянні:

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k).$$

Між рівняннями (1.1), (1.2) і (1.3) різниця полягає у детермінованості a_0 і $a_2 k$ у рівняннях (1.1) і (1.3) відповідно. Рівняння (1.1) являє собою модель без константи та тренду, ще має назву випадкового кроку, друге – має зміщення у вигляді константи a_0, а третє вміщує зміщення та детермінований лінійний часовий тренд.

У даних рівняннях об'єктом дослідження виступає γ. Якщо $\gamma = 0$, то послідовність $\{y(k)\}$ має одиничний корінь.

За тестом Дікі-Фуллера необхідно оцінити одне або ж більше, з рівнянь (1.1), (1.2) та (1.3), за допомогою МНК чи ММП, щоб проаналізувати параметр γ та його стандартну похибку. Ці дані допоможуть нам обчислити t −статистику, яку необхідно порівняти із значеннями в таблиці Дікі-Фуллера. Це порівняння стане ключовим і допоможе зробити висновок щодо нуль-гіпотези $\gamma = 0$.

Значення t −статистики різняться залежно від того, чи наявні в модулі зміщення a_0 та детермінований тренд $a_2 k$. Тому було визначено, що критичні значення для $\gamma = 0$ залежать від декількох характеристик, а саме: довжини вибірки та структури регресійного рівняння. Було введено наступні позначення: для рівняння (1.1) статистика має позначення τ, для (1.2) – τ_μ, і для (1.3) – τ_τ.

Якщо ж в моделі буде включена константа зміщення, проте $a_2 = 0$, тоді використовувати потрібно іншу частину таблиці критичних значень для t −статитики.

Якщо замінити рівняння (1.1), (1.2) та (1.3) на наступні рівняння авторегресії:

$$\Delta y(k) = \gamma y(k-1) + \sum_{i=2}^{p} \beta_i \Delta y(k-i+1) + \varepsilon(k), \quad (1.4)$$

$$\Delta y(k) = a_0 + \gamma y(k-1) + \sum_{i=2}^{p} \beta_i \Delta y(k-i+1) + \varepsilon(k), \quad (1.5)$$

$$\Delta y(k) = a_0 + \gamma y(k-1) + a_2 k + \sum_{i=2}^{p} \beta_i \Delta y(k-i+1) + \varepsilon(k), \quad (1.6)$$

то критичні значення залишаться незмінними.

Дікі і Фуллер запропонували наступні три статистики для перевірки об'єднаних гіпотез: $F-$ статистики ϕ_1, ϕ_2, ϕ_3 .

Тобто для об'єднаних гіпотез статистика ϕ_1 використовується для рівнянь (1.2), (1.5) щоб перевірити нуль-гіпотезу $\gamma = a_0 = 0$, відповідно статистика ϕ_2, для перевірки $\gamma = a_0 = a_2 = 0$ в рівняннях (1.3), (1.6) та статистика ϕ_3 для перевірки об'єднаної гіпотези $\gamma = a_2 = 0$.

Формула для обчислень статистик ϕ_1, ϕ_2, ϕ_3 :

$$\phi_i = \frac{[RSS_1 - RSS_2]r}{RSS_2(N-n)}$$

де n – число параметрів, оцінених для необмеженої моделі;

RSS_1 і RSS_2 – суми квадратів похибок, пораховані для моделей з обмеженнями та для моделей без обмежень;

N – число використаних спостережень;

r – число обмежень.

Завдяки обчисленню $F-$ статистик, та порівнянню їх із значеннями з таблиці Дікі-Фуллера, маємо можливість отримати рівень значущості, при якому обмеження на модель будуть давати вагомий результат.

Тоді генерування даних моделлю з обмеженнями буде сприйнято за нуль-гіпотезу, а без обмежень – за альтернативу. Якщо ж обмеження незначні, то сума квадратів похибок для моделі з обмеженнями буде наближатись до СКП для моделі без обмежень.

Звідси випливає, що значення ϕ_i буде незначним, а отже, при великих значеннях ϕ_i обмеження будуть грати суттєву роль і нуль-гіпотеза буде відхилена. Також необхідно перевірити гіпотези щодо наявності часового тренду, перевіривши значущості константи a_0 та коефіцієнта a_2.

При визначенні нуль-гіпотези як $\gamma = 0$, необхідно використовувати статистику τ_{β_τ} . Виконується перевірка чи $a_0 = 0$ при $\gamma = 0$. Якщо модель виду (1.6), то необхідно скористатись статистико τ_{α_τ} , якщо модель виду (1.5) – τ_{α_η} .

1.3.2 Тести на наявність гетероскедастичності

Для перевірки наявності гетероскедастичності можна використовувати наступні тести, описані у роботі [3], кожен з них в тій чи іншій мірі показує наявність цієї характеристики.

Почнемо з тесту *Бройша-Пагана/Годфрі*, і для початку запишемо рівняння наступної лінійної регресії, яку будемо перевіряти на наявність гетероскедастичності:

$$y(k) = \boldsymbol{x}^T(k)\beta + \varepsilon(k),$$

Де $\boldsymbol{x}^T(k) = [1\ x_2(k)\dots\ x_r(k)]$

Будемо вважати, що форма гетероскедастичності має вигляд:

$$E[\varepsilon(k)] = 0, \forall k,$$

$$var[\varepsilon(k)] = E[\varepsilon^2(k)] = \sigma_\varepsilon^2 = h\left(\alpha \boldsymbol{z}^T(k)\right),$$

де $\boldsymbol{z}^T(k) = [1\ z_2(k)\dots\ z_p(k)]$ – вектор відомих змінних;

h - невизначена функція що набуває тільки додатних значень;

$\alpha = [\alpha_1 \alpha_2 \dots\ \alpha_p]$ – невідомі коефіцієнти.

Тоді вигляд нуль-гіпотези буде наступним:

H_0: $\alpha_1 = \alpha_2 = \alpha_p = 0$, звідки випливає, що $\sigma_\varepsilon^2 = h(\alpha_1) = const$.

Для оцінки коефіцієнтів моделі такої нуль-гіпотези можна скористатись МНК, але вважаючи нормальним розподіл збурень правої сторони рівняння. Щоб перевірити гетероскедастичність, необхідно в першу чергу оцінити параметри вхідного рівняння лінійної регресії, за допомогою звичайного МНК, спочатку формуємо масив залишків $e(k) = y(k) - x^T(k)\beta$ та обчислюємо для них дисперсію $\sigma_\varepsilon^2 = \sigma_e^2 = N^{-1}\sum e^2(k)$.

Наступним кроком є знаходження оцінки регресії $\frac{e^2(k)}{\sigma_\varepsilon^2}$ на $z(k)$, використовуючи ЗМНК, і значень похибки $EES = \beta^T X^T X\beta - N\mu_y^2$, в якості μ_y виступає середнє значення послідовності $\{y(k)\}$.

Далі, приймаємо нуль гіпотезу $H_0\ \frac{1}{2}\ ESS \leftrightarrow \chi^2(p-1)$, звідси випливає, що при перевищенні значення ESS/2 над критичним значенням, вибраним з

розподілу, гіпотеза стосовно наявності гетероскедастичності буде непідтвердженою.

Для того, щоб застосувати даний тест для перевірки гетероскедастичності, необхідно мати значення змінних z. Зазвичай підходящі значення беруться з вектора регресорів *x(k)*.

Ще однією перевіркою є тест *Уайта*. Почнемо з побудови додаткової моделі регресії квадратів залишків, що згенерувались після застосування МНК до часового ряду. Вона містить в собі крім константи ще й всі ненадлишкові регресори, а також їх квадрати та добутки, і має наступний вигляд:

$$y(k) = a_0 + a_1 x_1(k) + a_2 x_2(k) + \varepsilon(k),$$

Дана модель допоможе виявити наявність гетероскедастичності. Дослідивши її детальніше, наважко помітити, що вектор вимірів незалежних змінних має наступний вигляд $[1\ x_1\ x_2]^T$, звідси випливає, що шукана множина регресорів виглядає наступним чином $[1\ \ x_1\ \ x_2\ \ x_1^2\ \ x_2^2\ \ \ x_1 x_2]$.

Знаючи залежність $NR^2 \leftrightarrow \chi^2(q)$, де q – кількість регресорів, виключаючи константу, R^2 коефіцієнт множинної регресії, а χ^2 відомий нам розподіл, з ступенем свободи, що може набувати велике значення, можна перевірити наявність гетероскедастичності, але не зрозуміти її форму та алгоритм оцінки параметрів. Тому додатково потрібно використовувати МНК

І в завершенні розглянемо тест *Голдфельда-Квандта*. Застосування даного тесту відбувається у випадках, коли є лише одна змінна, наявність якої забезпечує гетероскедастичність. Якщо показник σ_ε^2 характеризується додатною корельованістю з і-тим регресором x_i, то для проведення тесту необхідно спочатку проранжувати масив x_i, та видалити з аналізу *c* середніх значень. Після

чого обрахувати похибки RSS_1 та RSS_2, для окремо побудованих перших та останніх *(N — c)/2* значень регресій.

Припустимо що є гетероскедастичність, тоді якщо відношення $R = \frac{RSS_2}{RSS_1}$ матиме $F -$ розподіл із ступенем свободи $\frac{N-c-2r}{2}$. При альтернативній гіпотезі F буде мати велике значення.

Вагомим фактором, від якого залежить якість тестування є кількість середніх спостережень, що будуть видалені з аналізу. Якість тестування може погіршитись, при недостатньому, або занадто великому c. Для отримання найкращих результатів, прийнято брати $c \approx N/3$.

1.4 Деякі комп'ютерні системи для побудови моделей демографічних процесів

Практичне застосування методів прогнозування важко уявити без сучасних потужних комп'ютерних програм. Проте складним є вибір потрібної програми для побудови моделей демографічних процесів. З одного боку – бажання мати в наявності таку систему, що забезпечила б можливість обробки інформації довільної форми та довжини, із сотнями різних запрограмованих методів аналізу, з іншого – необхідність звести складність роботи з системою до нуля.

Сьогодні ж одною з кращих та найпотужніших готових систем для аналізу та прогнозування часових рядів є статистичний пакет Eviews, тому розглянемо характеристику його роботи.

Перед тим, як почати роботу, потрібно задати формат даних, після чого створити новий об'єкт, що має тип ряд, вказати кількість змінних і спостережень.

Eviews передбачає роботу з наступними типами даних: піврічні, річні, місячні, квартальні, тижневі, щоденні та недатовані, що є дуже зручним у використанні.

Для простоти введення та опису даних, краще імпортувати їх з інших Windows-додатків. Eviews підтримує наступні розширення файлів з даними: *.accdb, *.dht, *.ods, *.xls, *.html, *.saas.

Для легкої візуалізації вхідних та вихідних даних, в Eviews доступна можливість автоматичної побудови та відображення різних графіків, діаграм, корелограм, тощо. Більш того, ці графіки можна переглядати в одному вікні, не перемикаючись кожного разу між великою кількістю вікон в панелі меню, що значно економить час та енергію користувача. Всі числові характеристики, які нас цікавлять, можна переглянути за допомогою команд меню, без створення додаткових процедур.

Найголовніше, що за допомогою пакету Eviews, можна досліджувати задачі регресійного аналізу, аналізу часових рядів, моделювання стаціонарних та нестаціонарних процесів, тощо. У даному додатку передбачена робота з лінійною регресією, нелінійною регресією та моделями з дискретною залежною змінною.

A. Регресійний аналіз

a) Лінійна регресія

Eviews надає можливість провести оцінку коефіцієнтів лінійної регресії (статистика Фішера оцінки значимості регресії, статистика Дарбіна-Уотсона, коефіцієнт детермінації, коефіцієнт кореляції/кореляційна матриця) за допомогою методу найменших квадратів, максимальної правдоподібності, зваженим і нелінійним методами найменших квадратів. Також можна відразу провести поправку на гетероскедастичність, враховуючи характер залежності помилок від незалежної змінної. Визначення гетероскедастичності можливе методом Глейзера, Парку, а також воно доступне прямо за командою меню.

b) Нелінійна регресія

Для обрахунку нелінійної регресії, можна скористатись запропонованими наступними видами залежності: регресія експоненціального зростання, регресія логіт/пробіт, кусково-лінійна регресія, тощо. Щоб оцінити коефіцієнти нелінійної регресії можна скористатись нелінійним МНК і зваженим МНК.

Також присутній тест Боксу-Коксу, за допомогою якого можна підібрати вигляд гладкої функції, що якнайкраще відображає залежність.

c) Моделі з дискретною залежною змінною

В додатку Eviews наявна можливість будувати моделі як бінарного, так і множинного вибору, з порядковими або ж неврегульованими альтернативами. Після вибору необхідного методу, відповідно шуканій моделі отримаємо вихідні параметри: логарифмічні функції правдоподібності, коефіцієнт детермінації тощо.

B. Аналіз часових рядів

Для аналізу часових рядів у програмі Eviews передбачені наступні методи досліджень: модель авторегресії проінтегрованого сковзкого середнього, експоненціальне згладжування і прогнозування, моделі інтервенції АРПСС, аналіз розподілених лагів, спектральний аналіз ряду, тощо. Також доступні різноманітні операції з рядом, такі як: виділення тренду, знаходження різниць різного порядку, алгебраїчні операції та інші.

Побудова адаптивних моделей у пакеті Eviews доволі складна. Щоб побудувати модель необхідно визначити чимало попередніх показників: тип моделі, метод оцінювання параметрів, адаптаційні коефіцієнти, критерії збіжності тощо.

Проте процедура пошуку та згладжування тренду за допомогою ковзного середнього досить проста та зручна у використанні. Більш того, крім стандартних методів пошуку тренду у вихідних даних: спектрального аналізу, корелограм

АКФ і ЧАКФ, також можна задати перевірку спеціальним текстом через командний рядок.

C. Моделювання стаціонарних процесів

Додаток Eviews надає ряд можливостей для моделювання стаціонарних процесів. Доступні тести на наявність стаціонарності: звичайний та розширений Діки-Фуллера, тест Люнга-Боксу для ідентифікації моделі АРКС, критерії Шварця та Акайке для селекції двох значимих моделей АРКС, які використовують дисперсію помилки, що є мінімальною.

Також доступна можливість побудови моделі стаціонарних рядів, з урахуванням зміни дисперсії, та тестування її на гетероскедастичну складову. Є важливим можливість перевірки коінтеграції між змінними, за допомогою процедури Йохансена.

D. Моделювання нестаціонарних процесів

Для моделювання нестаціонарних процесів, використовується модель АРІКС, де порядок інтеграції – це порядок взяття різниць. Перевірка моделі АРІКС використовує розширений тест Діки-Фуллера. Процес перевірки на стаціонарність є автоматизованим і відбувається як тільки буде взято різницю першого чи другого порядків.

Програмна система Eviews являє собою комплекс, який об'єднує в собі максимальні функціональні можливості з проведення статистичних та економетричних досліджень, демонструє здатність до розвитку та інтеграції з іншими програмними продуктами, в тому числі MS Office. Система буде корисна при проведенні фінансового аналізу, макроекономічному прогнозуванні, моделюванні, аналізі цін і складанні прогнозу продажів.

Можливості Eviews забезпечують проведення всебічного аналізу даних, проведення різних типів прогнозів, а також регресійного аналізу. З його допомогою можна досить швидко отримати більшість статистичних показників

на основі введених вихідних даних і потім спрогнозувати розвиток економічних процесів для конкретної проблемної області або завдання.

Проте з іншого боку стоїть нагромадженість не завжди потрібними функціями, не зовсім простий та зрозумілий інтерфейс, існування нових інструментів, що дозволять аналізувати дані врази швидше. Відсутність інтеграції з іншими операційними системи унеможливлює цілковиту доступність програми для різних користувачів. Тому постає питання про створення програмного продукту, який би вирішував вищезгадані незручності.

1.5 Постановка задачі дослідження і висновки до розділу

Після дослідження демографічних процесів в Україні та Німеччині, можна зробити висновок, що за більшістю з обраних критеріїв порівняння, Україна має гірші показники, ніж Німеччина, і лише за рівнем народжуваності держави показали максимально схожий результат. За останні п’ять років чисельність населення України зменшилась на 3 273 000 осіб. Від'ємне сальдо міграції, зниження тривалості життя, поширення соціальних хвороб стали характерними ознаками демографічних процесів з початку XXI століття і дають підстави кваліфікувати їх як ознаки демографічної кризи.

Було досліджено ряд моделей для створення чіткого прогнозу демографічних процесів, коротко описано декілька з них, а саме: моделі авторегресії, авторегресії з ковзним середнім, парної регресії, змішаної регресії, множинної регресії, регресії з нелінійними змінними, регресії з нелінійними параметрами, моделі для гетероскедастичних процесів,та байєсівські мережі.

Також розглянуто тест Дікі-Фуллера на перевірку стаціонарності та тести на наявність гетероскедастичності: Бройша-Пагана/Годфрі, Уайта та

Голдфельда-Квандта. Після аналізу однієї із найкращих готових комп'ютерних систем для побудови моделей – Eviews, стало зрозуміло, що пакет містить собі функції, які не завжди потрібні користувачу, має не зовсім простий та зрозумілий інтерфейс, зокрема не варто ігнорувати факт існування нових інструментів, що дозволять аналізувати дані врази швидше.

Тому необхідна система, яка б мала інтеграцію не лише з операційною системою Windows, щоб була гнучка і цілковита доступність програми для різних користувачів. Тому постає питання про створення програмного продукту, який би вирішував вищезгадані незручності, та за допомогою якого можна чітко спрогнозувати демографічні процеси на прикладі України.

РОЗДІЛ 2 ВИБІР ОПИС МАТЕМАТИЧНИХ МОДЕЛЕЙ ДЛЯ ДЕМОГРАФІЧНИХ ПРОЦЕСІВ

2.1 Математичні моделі для нелінійних нестаціонарних процесів

2.1.1 Моделювання процесів з трендом

А. Авторегресія з трендовою складовою:

$$y(k) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + \sum_{j=1}^{m} b_j k^j + \varepsilon(k),$$

Авторегресія з інтегрованим ковзним середнім:

$$y(k) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + \sum_{j=1}^{q} b_j v(k-j) + \varepsilon(k),$$

Якщо серед коренів характеристичного рівняння для обраної авторегресійної частини є одиночний, то дані формули описують тренд і, накладені на нього коливання.

В. Сплайни для опису квадратичних, кубічних і трендів вищих порядків

Сплайн, що складається із поліномів першого порядку називається лінійним. На жаль, точність інтерполяції, отриманої лінійними сплайнами є невисокою. Однак в деяких випадках кусочно-лінійна апроксимація функції може виявитися кращою, ніж апроксимація вищого порядку. Сплайном першого порядку називається неперервна на відрізку $[a, b]$ лінійна на кожному частковому відрізку функція $f(x)$. Його позначають $S_1(x)$.

Інтерполяційним для даної функції $f(x)$ називається сплайн, який

задовольняє умовам $S_1(x_i) = y_i, i = 0, \dots, m$. Графік лінійного інтерполяційного сплайну $S_1(x)$ – це ламана, яка проходить через задані точки.

Нехай $x \in [x_i, x_{i+1}]$, формула для сплайну $S_1(x)$ на такому проміжку виглядає наступним чином:

$$S_1(x) = y_i \frac{x_{i+1} - x}{h_i} + y_{i+1} \frac{x - x_i}{h_i},$$

де $S_1(x)$ – сплайн першого порядку;

$y_i = f(x_i), i = 0, \dots, m, a = x_0 < \cdots < x_n = b$ – задані значення;

$h_i = x_{i+1} - x_i$.

Для апроксимації даних часто використовують сплайн Ерміта – це сплайн третього порядку, похідна якого приймає у вузлах сплайна задані значення. У кожному вузлі сплайна Ерміта задано не тільки значення функції, а й значення її першої похідної. Сплайн Ерміта має безперервну похідну, але друга похідна у нього розривна. Точність інтерполяції значно краще, ніж у лінійного сплайна.

Розглянемо кубічний сплайн. Деяка функція $f(x)$ задається на проміжку $[a, b]$, який розбито на частини $[x_{i-1}, x_i], a = x_0 < \cdots < x_n = b$. Кубічним сплайном називається функція $S(x)$, яка на кожному з відрізків $[x_{i-1}, x_i]$ є многочленом не вище третього порядку; має неперервну першу та другу похідні на всьому відрізку $[a, b]$; в точках x_i виконується рівність $S(x_i) = f(x_i)$.

Така функція має вигляд

$$S_i(x) = a_i + b_i(x - x_i) + \frac{c_i}{2}(x - x_i)^2 + \frac{d_i}{6}(x - x_i)^3,$$

де $S_i(x)$ – кубічний сплайн;

$$a_i = S_i(x_i) = f(x_i);$$

$$b_i = S_i'(x_i) = \frac{f_i - f_{i-1}}{h_i} - \frac{h_i(2c_i + c_{i-1})}{6};$$

$$c_i = S_i''(x_i);$$

$$d_i = \frac{c_i - c_{i-1}}{h_i}.$$

C. Експоненційна модель

Експоненційним трендом називають тренд, який записується наступним рівнянням:

$$y(k) = ap^k + \varepsilon(k),$$

де $y(k)$ – основна змінна процесу;

a – вирівняний тренд,

p – темп зміни рівня. Якщо $p > 1$ тренд із прискоренням, що зростає. Якщо $p < 1$, скорочення рівнів та зниження амплітуди.

Експонентний тренд притаманний процесам, для яких непередбачені обмеження на зростання рівню. Це означає, що на практиці він може розвиватися тільки на обмеженому часовому періоді

$$y(k) = \exp(\ln(a) + \ln(pk)) + \varepsilon(k).$$

D. Модель з циклічним компонентом

Для опису нерегулярних або ж відносно довгих в часі періодів зростання чи спаду рівнів часового ряду, застосовується циклічний компонент. Для ідентифікації даного компоненту необхідно мати додаткову інформацію про обраний процес.

Моделі з циклічним компонентом зазвичай використовуються для прогнозів зміни показників внаслідок спадів чи падінь в сфері економіки, зміни динаміки популяцій під час різних захворювань та епідемій тощо.

Тренд у даному випадку описують комбінацією тригонометричних функцій:

$$y(k) = a_0 + \sum_{i=1}^{p} a_i sin(\omega_0 i) + \sum_{j=1}^{q} b_j \, cos(\omega_0 i) + \varepsilon(k),$$

де $y_n(k)$ – основна (залежна) змінна;

$a_0, \{a_i, b_i\}, i = 1, \dots, n$ – коефіцієнти моделі;

Можна використовувати будь-яку детерміновану функцію, для опису тренду, головне наявність її відповідності характеру його зміни у часі.

E. Поліноміальна модель

Найбільш гнучкою та адаптивною з точки зору апроксимативних можливостей є поліноміальна модель, що має наступний вигляд:

$$y(k) = a_0 + a_1 k + a_2 k^2 + \cdots + a_m k^m + \varepsilon(k),$$

де $y(k)$ – основна (залежна) змінна процесу;

$a_i, i = 0, \dots m$ – коефіцієнти моделі;

$\varepsilon(k)$ – випадковий процес, через вплив зовнішніх збурень

де час є тим визначальним фактором, що впливає на незалежну змінну *y* та входить до моделі у вигляді регресорів з різними степенями.

Використання дискретного часу, що збігається із натуральним рядом чисел, визначає цей показник як фіктивну змінну, що задається в моделі не маючи ніякого відношення до обліку часу.

F. Показникові, степеневі, обернені та інші моделі

Вибір математичної моделі у більшій мірі залежить від сфери застосування. Найбільш поширеними базовими трендовими моделями є показникові, степеневі та обернені, що мають наступний вигляд:

$$y = ab^t,\ \ y = at^b,\ \ y = ae^{bt},\ \ y = a + b\frac{1}{t},\quad y = \frac{1}{ab^t + c},$$

$$y = \frac{a}{1 + ae^{-bt}},\ \ y = \frac{c}{a + bt},\ldots .$$

2.1.2 Моделювання гетероскедастичних процесів

А. Модель АРУГ

Моделі, для яких дисперсія є змінною в часі, називаються гетероскедастичними. Зазвичай, для опису змінної дисперсії застосовують модель типу АР(q) до оцінок залишків в квадраті. Тоді вони приймають наступну форму:

$$\widehat{\varepsilon^2}(k) = \alpha_0 + \alpha_1\widehat{\varepsilon^2}(k-1) + .. + \alpha_q\widehat{\varepsilon^2}(k-q) + v(k), \quad (2.1.)$$

де $v(k)$ – білий шум.

Якщо трішки інтерпретувати вигляд даного рівняння, то його можна використовувати для прогнозу умовної дисперсії на один крок:

$$E_k\left[\widehat{\varepsilon^2}(k+1)\right] = \alpha_0 + \alpha_1\widehat{\varepsilon^2}(k-1)+..+\alpha_q\widehat{\varepsilon^2}(k+1-q).$$

Рівняння (2.1) дістало назву АРУГ, тобто авторегресійним умовно гетероскедастичним. Зовнішні збурення можуть бути представлені в наступній мультиплікативній формі:

$$\varepsilon^2(k) = v^2(k)[\alpha_0 + \alpha_1\varepsilon^2(k-1)],$$

де $v(k) -$ білий шум, в основі якого мультиплікативне збурення, і разом з $\varepsilon(k-1)$ це незалежні статистичні величини.

З всього вищесказаного можна зробити висновок що умовне середнє $\varepsilon(k)$ можна знайти за наступною формулою:

$$E[\varepsilon(k)|(\varepsilon(k-1),\varepsilon(k-2),\dots] = E[v(k)]E[\alpha_0 + \alpha_1\varepsilon^2(k-1)]^{\frac{1}{2}} = 0.$$

Звідси знайдемо безумовну дисперсію $\varepsilon(k)$, за умови, що $\sigma_v^2 = 1$

$$E[\varepsilon^2(k)|(\varepsilon(k-1),\varepsilon(k-2),\dots] = \alpha_0 + \alpha_1\varepsilon^2(k-1).$$

Для того щоб безумовна дисперсія задовольняла обов'язкову умови додатності, необхідно, щоб і α_0, α_1 також були додатними величинами. А для забезпечення стійкості АР, α_1 повинно лежати у проміжку (0;1).

Безумовна дисперсія також може мати наступний вигляд:

$$E[\varepsilon^2(k)] = \frac{\alpha_0}{1-\alpha_1}$$

Оскільки процес $\varepsilon(k)$ не впливає ні на безумовне середнє, ні на безумовну дисперсію, то по аналогії можна визначити, що умовне середнє рівняється нулю. Якщо виконуються умови: $v(k)$ і $\varepsilon(k-1)$ незалежні величини і $E[v(k)] = 0$, то умовне середнє матиме наступний вигляд:

$$E_{k-1}[\varepsilon(k)|\varepsilon(k-1), \varepsilon(k-2), \dots] = E_{k-1}[v(k)]E_{k-1}[\alpha_0 + \alpha_1\varepsilon^2(k-1)]^{\frac{1}{2}} = 0$$

А умовна дисперсія матиме наступний вигляд:

$$E_{k-1}[\varepsilon^2(k)|\varepsilon(k-1), \varepsilon(k-2), \dots] = \alpha_0 + \alpha_1\varepsilon^2(k-1).$$

Очевидно що умовна дисперсія $\varepsilon(k)$ залежить від значення $\varepsilon^2(k-1)$. Із ростом останньої зростатиме значення умовної дисперсії.

Структура даної моделі особлива тим, що значення умовного та безумовного середнього чисельно рівне нулю, що притаманно для показника адекватності. До того ж, послідовність $\varepsilon(k)$ є послідовно некорельованою.

Дану модель можна розширювати до довільного порядку:

$$\varepsilon(k) = v(k)\left(\alpha_0 + \sum_{i=1}^{q}\alpha_1\varepsilon^2(k-i)\right)^{\frac{1}{2}}.$$

В. Модель УАРУГ

Далі розглянемо ще одну модель гетероскедастичного процесу УАРУГ(p,q).

Узагальнена модель АРУГ складається із авторегресії та ковзного середнього відносно дисперсії процесу. Для даної моделі, білий шум, себто зовнішнє збурення, що діє на процес $\{y(k)\}$, являється процесом АРКС. Тому цілком передбачувано, гетероскедастичному процесу будуть відповідати похибки моделі АРКС.

Це твердження доведене наступним чином: зробимо припущення про належність $\{y(k)\}$ до процесу АРКС, якщо підтверджена адекватність даної моделі до обраних процесів, то АКФ і ЧАКФ залишків вказуватимуть на процес білого шуму. Для визначення порядку процесу УАРУГ, слід використовувати АКФ квадратів залишків. Запишемо наступне рівняння:

$$E_{k-1}[\varepsilon^2(k)] = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon^2(k-i) + \sum_{i=1}^{p} \beta_i h(k-i).$$

Умовне і безумовне середнє для $\varepsilon(k)$ дорівнюватиме нулю, якщо процес $\{v(k)\}$ є некорельованим із значенням $\varepsilon(k-i)$, білим шумом. Тоді безумовне математичне сподівання обчислюватиметься за наступною формулою:

$$E[\varepsilon(k)] = E\left[v(k)(h(k))^{1/2}\right] = 0$$

Вибіркова умовна дисперсія має ж наближено наступний вигляд:

$$h_b(k) = \frac{1}{k-1}\sum_{i=1}^{k}[y(i) - \overline{y_b}(k)]^2,$$

де $\overline{y_b}(k)$ – умовне середнє.

Для забезпечення скінченності умовної дисперсії, корені характеристичного рівняння мають лежати всередині кола на комплексній площині одиничного радіусу.

С. Модель ЕУАРУГ

Наступною для розгляду візьмемо модель ЕУАРУГ, тобто експоненціальна модель УАРУГ. Дана модель виключає можливі недоліки моделей АРУГ та УАРУГ. Розглянемо логарифм умовної дисперсії, що визначається за допомогою функції нормальних похибок, і має наступний вигляд:

$$\log[h(k)] = c_0 + \sum_{i=1}^{\infty} c_1 g[y(k-i)],$$
$$g(y) = \alpha y(k) + \beta[|y(k)| - E|y(k)|],$$

де $E[g(y)] = 0$, що означає визначення $\log[h(k)]$ як лінійної фільтрації, а його рівняння як ковзного середнього нескінченного порядку.

На практиці ж застосовують представлення, що має наступний вигляд:

$$\log[h(k)] = c_0 + \frac{1+\alpha(L)}{1-\beta(L)} g[y(k)],$$

за умови, що $1+\alpha(L)$ та $1-\beta(L)$ не матимуть спільних коренів.

Значення $\beta[|y(k)| - E|y(k)|]$ відповідає за ефект абсолютного значення.

Ще одна інтерпретація моделі ЕУАРУГ записується наступним чином:

$$\log[h(k)] = \alpha_0 + \sum_{i=1}^{p} \alpha_1 \frac{|\varepsilon(k-i)|}{\sqrt{h(k-i)}} + \sum_{i=1}^{p} \beta_1 \frac{|\varepsilon(k-i)|}{\sqrt{h(k-i)}} + \sum_{i=1}^{q} \gamma_1 \log[h(k-i)] + v(k).$$

2.2 Методика побудови моделей інтегрованих та гетероскедастичних процесів

Для того щоб побудувати математичну модель обраного процесу, необхідно дотримуватись наступної послідовності кроків:

А. Проведення повного аналізу процесу.

Будь-яка система адаптивного прогнозування бере початок з визначення процесу, для якого буде побудована модель, аналізу існуючих підходів прогнозування, тощо.

Для проведення якісного аналізу зазвичай досліджують спеціальну літературу, у якій можна знайти особливості поведінки того чи іншого процесу, порівняльну характеристику моделей чи критеріїв оцінки її параметрів. На даному етапі також обирається тип та структура моделі, оскільки ці дані грають важливу роль у подальшому створені якісного прогнозу та розробки керуючої системи. А для того, щоб прогноз був якісним побудована модель повинна мати високий показник адекватності та враховувати поведінку реальних процесів.

Етап аналізу є надзвичайно важливим, адже це основа, завдяки якій розширюється кругозір особи, що створює модель. Він також спрямований на дослідження вимірності прогнозу, забезпечення логічних зв’язків між вхідними змінними, зовнішній вплив та класифікація його типу, встановлення можливості розбиття даного процесу на окремі частини, що спростить керування в цілому, визначення функціональності кожного рівня зв’язків та їх місце у загальній ієрархії процесу, і найголовніше – дослідження переваг та недоліків розробки моделі та її адаптивність у подальшому використанні.

Отож, після детального аналізу отримана інформація спрямовується на попередню оцінку структури моделі.

В. Попередня обробка даних

На жаль, рідко можна знайти готову ідеальну модель для використання. Тому у більшості випадків проводиться корегування структури та параметрів вже існуючої обраної моделі. Якщо ж це не можливо, то з'являється необхідність у побудові нової, на визначених статистичних даних. Якість цих даних має велику цінність, тому при їх виборі необхідно враховувати їх коректність, інформаційну наповненість та синхронність. Для цього і потрібен етап попередньої обробки даних, що забезпечить приведення їх до потрібної форми, і забезпечить можливість подальшого коректного застосування при використанні методів оцінок параметрів моделі.

Першим кроком є очищення моделі, це означає, що дані не мають мати дублікатів, аномалій, шумів, протиріч чи фіктивних значень, що приводили б до зниження їх якості та ефективності роботи алгоритму аналізу. Далі йде оптимізація даних, тобто визначення незначимих ознак та їх вилучення, зменшення розмірності тощо. При оптимізації дані адаптуються та збільшують ефективність аналізу конкретної задачі.

До основних операцій, що застосовують на етапі обробки даних належать: нормування та корегування даних, а також обчислення перших різниць чи різниць вищих порядків для аналізу часового ряду.

С. Аналіз наявності нелінійностей

Визначення наявності нелінійностей та класифікація їхнього типу в досліджуваному процесі є доволі складною процедурою при формуванні структури моделі.

Зазвичай проводять візуальний аналіз отриманих даних та формальні тести, що показують наявність нелінійностей. Якщо в особи, що формує модель є багатий досвід, то для неї не складе труднощів візуально виявити фрагменти, що показують ознаки нелінійності та гетероскедастичності, які мають вагомий вплив на якість моделі.

Проте передбачені інші шляхи вирішення цієї проблеми. Користуються популярністю використання формальних тестів, вибіркових кореляційних функцій, що розраховуються за вибірками статистичних даних, також використання методу групового урахування аргументів, тощо.

Деякі тести на наявність тренду та гетероскедастичності детально представлені у першому розділі даної роботи.

D. Формування інших елементів структури моделі

Після підготовки та попередньої обробки даних, постає необхідність для оцінки структури та параметрів математичних моделей інших моделей, потенційних-кандидатів, обраних на для процесу прогнозування. Ключовим моментом являється правильний вибір структури моделі, яка включатиме в собі наступні елементи:

– Кількість рівнянь, що лежать в основі моделі, або так звана вимірність моделі.
– Максимальний порядок різницевих або диференційних рівнянь, що складають модель.
– Нелінійність моделі та її тип.
– Оцінка часу, що витрачається при затримці реакції відносно входу.
– Тим зовнішніх збурювань процесу.

Зазвичай для одного процесу підбирають декілька моделей, і на основі оцінки множини статистичних параметрів цих моделей, обирають одну – найкращу.

Після визначення вищезгаданих елементів, можна сформувати модель і продовжити її подальший аналіз.

E. Оцінка коефіцієнтів потенційних моделей

Для оцінки параметрів рівняння застосовують принцип збереження. В основі цього принципу лежить правило про не перевищення кількості

оцінюваних коефіцієнтів над кількістю необхідних для збереження в моделі основних статистичних характеристик процесу.

В процесі моделювання процесів, дотримуватись принципу апроксимації поведінки процесу за допомогою рівнянь, а не розпису їх до найдрібніших елементів, враховуючи факт можливості існування різних за своєю структурою моделей з однаковим набором властивостей.

Також необхідно щоб справедливим було виконання можливості генерації початкового ряду моделі на основі якого була проведена оцінка коефіцієнтів, шляхом інверсування отриманого рівняння. Це означатиме співпадіння моделі з досліджуваним процесом за такими характеристиками як дисперсія та коваріація.

Для оцінки параметрів моделі застосовуються такі методи як: метод найменших квадратів, метод максимальної правдоподібності, нелінійний метод найменших квадратів, метод допоміжної змінної та інші.

Після чого оцінені моделі-кандидати перевіряють на адекватність і з них обирається одна – найкраща.

2.3 Критерії якості (адекватності) математичних моделей і прогнозів

Не залежно від виду і способу побудови математичної моделі прогнозування, питання про доцільність її застосування для аналізу і прогнозування економічних та демографічних явищ і процесів, може бути вирішене лише після встановлення її адекватності, тобто відповідність цієї моделі досліджуваному процесу. Зрозуміло, що повної відповідності моделі до реального об’єкту дослідження бути не може, отже поняття адекватності є дещо

умовним. Тому адекватність застосовується не до моделі в цілому, а до її властивостей, які вважаються істотними для дослідження.

Для окремої оцінки коефіцієнтів математичної моделі в статистичному сенсі, визначення інтегральної похибки відносно вихідного часового ряду, перевірки наявності кореляції між значеннями похибок моделі, використовують наступні критерії перевірки адекватності моделі:

а) $t-$ статистика Стьюдента.

Для визначення значимості коефіцієнтів регресії, окремо для кожного застосовується $t-$ статистика, що обчислюється пакетами статистичної обробки даних. Для цього використовують наступну формулу:

$$t_a = \frac{a` - a_0}{SE_a},$$

де $a` -$ отримана за допомогою пакета оцінка коефіцієнта;

$SE_a -$ стандартна похибка оцінки коефіцієнта; $a_0 -$ нуль-гіпотеза.

Щоб отримати кращу оцінку коефіцієнту для моделі, необхідно мінімізувати значення стандартної похибки.

Від значення довжини вибірки N та кількості параметрів p, а також рівня їх значущості α, залежить загальна значущість оцінки коефіцієнта. Вибраний рівень значущості α означає, що при оцінюванні регресії, допустиме помилкове прийняття рішення не перевищує його значення. За допомогою вищезгаданих параметрів можливо обрати значення $t_{\text{крит}}$. Нуль-гіпотеза щодо незначущості досліджуваного коефіцієнту приймається, коли $t_a \in (-t_{\text{крит}};\ t_{\text{крит}})$, в інакшому випадку коефіцієнт є значущим, і нуль-гіпотеза не приймається. Для отримання вищої значущості кожного коефіцієнта, очевидним є потреба максимізації значення t_a.

b) Сума квадратів похибок моделі

Обчислюється сума квадратів похибок досліджуваної моделі за наступною формулою:

$$SSE = \sum_{k=1}^{N} [y`(k) - y(k)]^2 ,$$

Де $y`(k) = a`_0 + a`_1 y`(k-1) + a`_2 y`(k-2) + b`_1 x(k) + b_2 z(k)$;

$y(k)$ −вимірювання;

Вибирати слід ту модель, для якої сума квадратів похибок приймає найменше значення.

c) Коефіцієнт детермінації

Як відомо, дисперсію часового ряду часто використовують як міру інформативності. Відношення дисперсії основної змінної частини часового ряду, отриманої з рівняння, до вибіркової дисперсії цієї ж змінної називається коефіцієнтом детермінації R^2, що обчислюється за наступною формулою:

$$R^2 = \frac{var(y)`}{var(y)} .$$

Щоб модель вважалася адекватною, необхідно виконання наступної умови: $R^2 \rightarrow 1$.

d) Статистика Дарбіна-Уотсона

Статистика Дарбіна-Уотсона застосовується для визначення ступеня корельованості похибок моделі, і обчислюється за формулою:

$$DW = 2 - 2\rho,$$

де $\rho = cov[e(k)] = E[e(k)e(k-1)]$ – коефіцієнт кореляції між сусідніми значеннями випадкової змінної.

Чим більше значення параметра, тим менший показник корельованості, а при її відсутності, коефіцієнт приймає максимально прийняте значення, що чисельно дорівнює 2.

e) Інформаційний критерії Акайке

Інформаційний критерій Акайке залежить від наступних показників: кількості вимірів *N*, кількості параметрів, що оцінюються *p* та від суми квадратів похибок. Визначається за наступною формулою:

$$\mathrm{AIC} = N \ln\left[\sum_{k=1}^{N} e^2(k)\right] + 2p$$

Для того щоб отримати кращу модель, необхідно мінімізувати значення суми квадратів похибок. Даний критерій перегукується із СКП, але наявність інших, вищезгаданих показників, робить його більш широким та інформативним.

f) Коефіцієнт Тейла

Коефіцієнт Тейла є одним із найважливіших показників, що описує точність та сумісність досліджуваної моделі. Він обчислюється за формулою:

$$U = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y`_i)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i)^2} + \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y`_i)^2}}.$$

Область значень даного показника знаходиться в межах (0;1). Модель не годиться для прогнозування, при значенні U=1, у випадку коли значення U

наближається до нуля, прогнозовані ряди та реальні ряди збігаються, що означає ідеальність опису вхідних даних.

Також коефіцієнт Тейла можна розкласти на суму відношень упередженості U^M, коваріацій U^C та варіацій U^S. Що обчислюються наступним чином:

$$U^M = \frac{(y_i - y`_i)^2}{\frac{1}{N}\sum_{i=1}^{N}(y_i - y`_i)^2}$$

$$U^C = \frac{2(1-\rho)(\sigma_{actual} - \sigma_{fitted})^2}{\frac{1}{N}\sum_{i=1}^{N}(y_i - y`_i)^2}$$

$$U^S = \frac{(\sigma_{actual} - \sigma_{fitted})^2}{\frac{1}{N}\sum_{i=1}^{N}(y_i - y`_i)^2}$$

Відношення упередженості U^M показує чи модель завищує прогноз. Якщо більш точно, то визначає наявність систематичних відхилень часового ряду. Щоб отримати кращу модель, необхідно мінімізувати значення відношення упередженості. Якщо U^M дорівнює нулю, то модель є дуже якісною.

Для визначення кореляції між прогнозованими та реальними рядами використовується відношення коваріацій. При ідеальній кореляції вищезгаданих рядів значення U^C має бути рівним 0.

Щоб перевірити наявність у моделі достатньої динаміки для поглинання варіацій реальних рядів, слід використовувати U^S. Так само, як у випадку з відношенням упередженості, чим менше значення відношення варіацій тим якіснішою є модель.

Сума трьох зазначених статистик рівна 1, і визначає їхній взаємозв'язок.

g) Статистика Фішера

За допомогою статистики Фішера можна визначити ступінь адекватності в загальному. Для того, щоб оцінити адекватність досліджуваної моделі, необхідно перевірити виконання наступної умови:

$$F > F_{\text{крит}},$$

Значення $F_{\text{крит}}$ можна знайти в таблиці, аналогічним чином як для t-статистики, а значення F залежить від коефіцієнта детермінації. Звідси випливає, що із підвищенням значення статистики Фішера росте значення ступеня адекватності моделі.

Крім ступеня адекватності моделі, важливим показником є якість отриманого прогнозу. Оскільки отримані результати є випадковими величинами, для їхнього оцінювання слід використовувати деякі статистичні критерії. Отож, розглянемо деякі з них.

h) Дисперсія і стандартне відхилення прогнозу

На практиці трапляються випадки, коли сума похибок прогнозів є нульовою, через різні знаки значень похибок, саме тому було введено інші міри похибки. Для вимірювання ступеня розсіювання значень змінної навколо її середнього, використовують стандартне відхилення $\sigma(k)$, яке чисельно дорівнює квадратному кореню з дисперсії:

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(y(k) - y`(k))^2}\,,$$

де $y(k)$ – значення змінної по факту;

$y`(k) -$ прогнозоване значення.

Одним із основних характеристик якості прогнозу є показник стандартного відхилення залишків. Дана статистична характеристика є досить широко використовуваною в аналізі поведінки процесів. Вона виступає як міра ризику у фінансових процесах, міра розсіювання значень змінних стану, як характеристика надійності в задачах управління запасами, тощо.

Дисперсія ж характеризується квадратичною характеристикою, тому під час її обчислення не відбувається компенсація знаків, що вирішує проблему із накладанням сум похибок. Для більш глибокої точності прогнозу, є можливість перевірки дисперсії і стандартного відхилення на статистичну значимість.

Також, маючи відносно велику множину вхідних значень, можна зробити висновок, що прогнозоване значення, отримане на виході, точно лежатиме в інтервалі, що визначається як $\pm 2\sigma$.

і) Середнє абсолютне значення похибки

Для обчислення середнього абсолютного значення похибки необхідно скористатись виразом експоненціального середнього, після чого отримаємо наступний результат:

$$\text{САП}(k) = \alpha\,|e(k)| + (1-\alpha)\text{САП}(k-1),$$

де $e(k) -$ похибка прогнозу,

$\alpha \in (0;1)$, при цьому значенні ми завжди отримуємо додатній результат шуканого значення.

Для визначення САП можна скористатись ще одним виразом:

$$\text{САП} = \frac{1}{N}\sum_{k=1}^{N} |y(k) - y`(k,k)|^2.$$

Цікавим є те, що для класу статистичних розподілів випадкових величин вірним є твердження про те, що стандартне відхилення є більшим та головне пропорційним значенню САП. Для різних розподілів маємо дещо різні коефіцієнти пропорційності. Проте, це дозволяє зробити висновок про вигляд наступного рівняння: $\sigma(k) \approx 1{,}25\text{САП}$.

Отже, для оцінювання стандартної похибки прогнозу необхідно знайти фактичне значення похибки прогнозу, порахувати нове значення САП та знайти поточне стандартне відхилення.

j) Середній квадрат похибки і сума квадратів похибок

Якщо розглядається випадок де середній квадрат похибки визначається для випадку однокрокового прогнозування, то він обчислюється за наступною формулою:

$$\text{СКП} = \frac{1}{N}\sum_{k=1}^{N}[y(k) - y`(k)]^2.$$

Якщо ж прогнозування неоднокрокове, то необхідно скористатись наступним виразом для обчислення середнього квадрату похибки:

$$\text{СКП} = \frac{1}{s}\sum_{i=1}^{s}[y(k+i) - y`(k+i,k)]^2.$$

Не важко помітити збіг обох формул, проте за допомогою різних форм можна чітко відслідкувати тип прогнозу.

Щодо показника суми квадратів похибок, то у системах математичного моделювання прийнято використовувати наступну формулу для визначення статичних розрахунків:

$$\text{СмКП} = \sum_{k=1}^{N} [y(k) - y`(k)]^2.$$

СКП та СмКП вважаються найпопулярнішими критеріями у порівняльному аналізі, вони є корисними інтегральними показниками якості, при виборі найкращої моделі для обрахунків оцінок прогнозу.

k) Середня абсолютна похибка у відсотках

Не важко здогадатись, що середнє абсолютних значень похибок оцінок прогнозу, представлене у відсотках, і являє собою значення середньої абсолютної похибки у відсотках, що визначається за наступною формулою:

$$\text{САПП} = \frac{1}{N} \sum_{k=1}^{N} \frac{|e(k)|}{|y(k)|} \times 100\%$$

Аналогічно для випадку неоднокрокового прогнозу:

$$\text{САПП} = \frac{1}{s} \sum_{i=1}^{s} \frac{|e(k+i)|}{|y(k+i)|} \times 100\%$$

Дана характеристика широко використовується у порівняльному аналізі якості прогнозування процесів різними методами.

Слід пам’ятати, що при наближенні $y(k+i)$ або $y(k)$ до нуля, САПП, буде прямувати до нескінченності, і для можливості оцінки прогнозування в такому випадку, нульові значення пропускаються з відповідним врахуванням цього факту в параметрах s та N.

l) Середня похибка і середня похибка у відсотках

На скільки є зміщеними фактичні значення від прогнозованих допомагає визначити середня похибка. СП визначається за наступною формулою:

$$\text{СП} = \frac{1}{N}\sum_{k=1}^{N} e(k)$$

Або, знову ж таки для неоднокрокового прогнозу:

$$\text{СП} = \frac{1}{s}\sum_{i=1}^{s} [y(k+s) - y`(k+s,k)]$$

При похибках із різними знаками, СП буде зменшувати своє значення. Щоб обчислити середню похибку у відсотках, скористаємось наступними формулами:

$$\text{СПП} = \frac{1}{N}\sum_{k=1}^{N} \frac{[y(k) - y`(k)]}{y(k)} \times 100\%$$

та

$$\text{СПП} = \frac{1}{s}\sum_{i=1}^{s} \frac{[y(k+s) - y`(k+s,k)]}{y(k+s)} \times 100\%$$

За допомогою СПП також визначають зміщеність прогнозу. Для отримання ідеальних значень прогнозу, необхідно щоб обидва показники наближались до нуля, проте очевидно що у реальному житті неможливо досягти такого результату. Тому загальноприйнятим задовільним результатом є значення <5%.

m) Максимальна та мінімальна абсолютна похибка

Максимальна абсолютна похибка обчислюється за наступною формулою:

$$\text{МАП} = \max_k\{|y(k) - y`(k)|\}, 1 \le k \le N,$$

Або

$$\text{МАП} = \max_i\{|y(k+1) - y`(k+1,k)|, \dots, |y(k+s) - y`(k+s,k)|\},$$
$$1 \le i \le s.$$

Мінімальна абсолютна похибка обчислюється за наступною формулою:

$$\text{МіАП} = \min_k\{|y(k) - y`(k)|\}, 1 \le k \le N,$$

Або

$$\text{МіАП} = \min_i\{|y(k+1) - y`(k+1,k)|, \dots, |y(k+s) - y`(k+s,k)|\},$$
$$1 \le i \le s.$$

Якщо об’єктом нашого дослідження є максимальні та мінімальні можливі відхилення прогнозів від фактичних значень на заданому інтервалі, то дані характеристики стануть нам у пригоді.

Ще одним фактором, що визначає якість досліджуваної моделі є те, знак тренду, тобто уміння точно спрогнозувати зміни напрямку розвитку описуваного процесу. Більшість моделей можуть показувати гарний результат на відтворені ряду, але бути неточними у прогнозуванні тренду та циклів, або ж навпаки, мати меншу адекватність, проте давати кращі результати при оцінці динамічних властивостей. Тому важливо знайти щось середнє між якістю апроксимації ряду

та її динамічними характеристиками. На жаль, для такого аналізу ще не створено формальних тестів.

Ще одним, не менш важливим показником є реакція моделі до початкового стартового періоду. Дану характеристику називають чутливістю. Якщо результати, згенеровані моделлю, можна класифікувати як варіантні до початкових умов, то дана модель вважається якісною, в інакшому випадку можуть бути отримані нестаціонарності певного типу. Інформацію про інші показники якості можна знати у [11] та [15] роботах.

2.4. Висновки до розділу

Одним з основних рішень у задачах прогнозування є вибір моделей. Тому в даному розділі було розглянуто основні види моделей інтегрованих та гетероскедастичних процесів, серед яких: авторегресійна модель та авторегресійна модель з ковзним середнім, поліноміальна, експоненційна моделі, модель, що описує циклічну складову, модель на основі сплайнів, показникові, степеневі, обернені моделі, також моделі АРУГ, УАРУГ, ЕУАРУГ.

Визначено основну методику їхньої побудови, що складається з наступних етапів: аналіз даних, попередня обробка даних, аналіз наявності нелінійностей, формування інших елементів структури моделі, оцінка коефіцієнтів моделей-кандидатів та вибір найкращої моделі.

Також наведено деякі критерії якості математичних моделей і прогнозів для перевірки обраної моделі на адекватність. Всі ці пункти є ключовими для отримання якомога точного прогнозу.

РОЗДІЛ 3 РОЗРОБКА І ЗАСТОСУВАННЯ СИСТЕМИ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ ДО ПОБУДОВИ МОДЕЛЕЙ ДЕМОГРАФІЧНИХ ПРОЦЕСІВ

3.1 Вимоги до СППР

Для того, щоб отримати прогноз високої якості, необхідно дотримуватись технологій, що основуються на застосуванні динаміки методів структуризації задач, методів обробки статистичних даних, моделей, процесів та критеріїв їх оцінки.

Застосування такого підходу гарантує отримання на виході адекватний прогноз та прийняті рішення, що на ньому ґрунтуються.

Проте, для забезпечення обробки великої кількості інформації, необхідно мати зручний, потужний та новітній інструментарій. Розглянемо декілька основних вимог користувача, що необхідні для створення СППР:

– Гнучка та зрозуміла мова програмування, яка забезпечить оптимальний розподіл ресурсів та пам’яті, необхідних для виконання процесу роботи, що забезпечить мінімізацію часу обробки запитів користувача.

– Максимально простий та зрозумілий інтерфейс користувача, для введення попередніх даних, демонстрації процесу обробки даних, представлення кінцевих результатів, тощо.

– Можливість збереження отриманих результатів.

Система підтримки прийняття рішень була розроблена мовою програмування python, містить в собі програмні модулі, що відповідають за зчитування та обробку вхідних даних, розбиття їх на тренувальну та основну вибірки, проведення тренування, власне прогнозування, обчислення характеристик якості моделі та прогнозу, вивід результатів, а також можливість

їх збереження та представлення часу, для обчислення ефективності роботи програми.

Для можливості використання розробленої програми необхідно бути забезпеченим наступними характеристики ПК:

– Частота процесора 1.6 -3.4 ГГц;
– Об’єм оперативної пам’яті 8 ГБ і більше;
– Інтегрований графічний адаптер;
– Операційна система: Windows 10, Linux;
– Роздільна здатність екрану: (1920x1080) Full HD;
– Розширення, що дають можливість переглядати файли формату *.csv, *.xlsx.

3.2 Розробка архітектури і функціональної схеми програми

Для проектування архітектури СППР важливим фактором є вибір підходу. Існують різні види підходів:

- Інформаційний

Передбачує належність СППР до класу інформаційних систем, створених для поліпшення характеру діяльності управляння, шляхом застосування засобів інформаційних технологій. Основними компонентами таких систем є інтерфейс «користувач» - «система», база даних, а також база моделей.

- Підхід заснований на знаннях

Полягає в об’єднанні технологій підтримки рішень та штучного інтелекту. Основна відмінність даного підходу полягає у спроможності системи зрозуміти проблему, зреагувати на запит користувача, обробити його та видати результат;

Структура даної СППР вміщує в собі наступні ключові елементи: мовна система, система знань, система обробки.

- Інструментальний підхід

Залежно від обраного від обраного напрямку, розглядаються наступні три рівні системи підтримки прийняття рішень: спеціалізовані СППР, генератори СППР, інструментарій СППР. Управління інтерфейсом такої СППР має реалізовуватись за допомогою меню, мови команд та мови запитань-відповідей. Аналіз даних контролюється в базі моделей, а управління представленнями підтримує різноманітні образи користувача стосовно проблеми, яку необхідно вирішити.

Після аналізу вищевказаних пунктів, за основу було обрано інформаційний підхід, оскільки він найбільш зрозумілий для користувача, і не використовує додаткових ресурсів та пам’яті, які можна використати власне при обробці статистичних даних та реалізації прогнозу.

Після вибору підходу проектування постає питання про вибір типу архітектури СППР. Розрізняють наступні типи архітектур:

- Текстово-орієнтовані

Застарілий вид архітектури, що містить мовну систему, базу знань, систему обробки даних, генерування результатів та їх представлення. Властивістю такої архітектури є гіпертекстова підтримка, що встановлює зв’язок між знаннями, які містяться у текстових файлах.

- Орієнтовані на бази даних

Архітектура СППР орієнтована на роботу з реляційними базами даних, що забезпечує обробку строго структурованих знань.

- Орієнтовані на використання електронних таблиць

Даний вид архітектури представляє основу процедурних знань, що служать командами для системи обробки даних та генерування результатів. Тип

архітектури призначений для виконання аналізу «що, якщо», для порівняння результатів дій над даними в таблицях. Іншими словами така архітектура прискорює порівняння всіх альтернатив.

- Основані на правилах

Особливість полягає у збереженні правил та опису станів до бази знань даних, та можливості формування висновку на основі вказаних правил. Окрім стандартних запитів, архітектура СППР дозволяє на основі правил створювати запити стосовно прийняття рішень (поради та пояснення деяких фактів).

- Гібридні

Очевидно, що така архітектура є універсальнішою та дає набагато більше можливостей для процесу обробки інформації та видачі результатів. Гібрідна архітектура є гнучкою, що спрощує інтеграцію СППР в майбутньому.

Саме гібридний тип архітектури був обраний для створення СППР, і він має наступний узагальнений вигляд, зображений на рисунку 3.1.

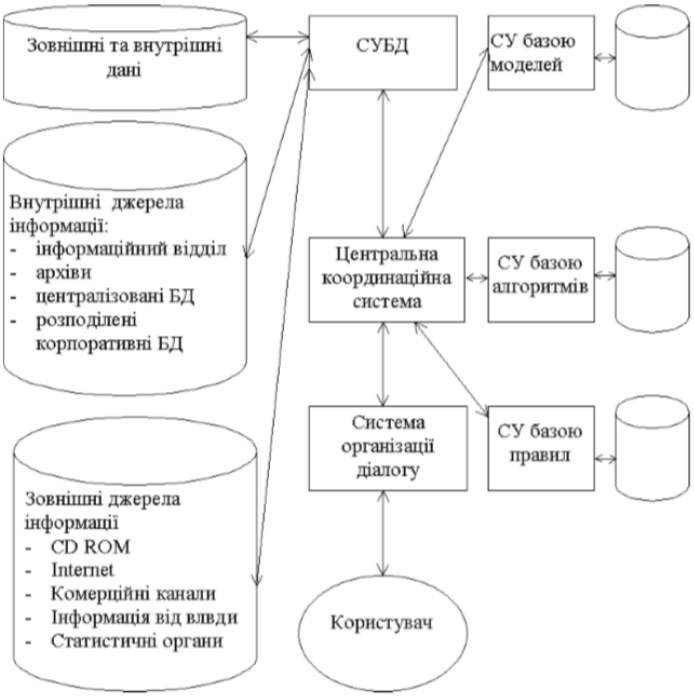


Рисунок 3.1 – Узагальнена архітектура гібридної СППР

Ще однією складовою при створенні СППР є функціональна схема програми. У розробленій СППР присутні наступні модулі, що реалізовують всі необхідні функції для коректної роботи програми:

- Config.py – модуль, що реалізовує зручну роботу із файлами.
- Metric.py – модуль з метрикою, що використовується при обчисленнях.
- Preprocess.py – модуль, що готує необхідні дані для прогнозу.
- Train_test.py – модуль, де формуються тестова вибірка і йде процес навчання моделі.
- Time.py – модуль, призначений для визначення часу роботи програми, за для аналізу ефективності вцілому.
- Predict.py – модуль, у якому власне вибирається країна та запускається процес прогнозування, виведення та запису результатів.

Також, у ході роботи було використано наступні бібліотеки:

a). numpy;

b). pandas;

c). xlrd;

d). lightgbm;

e). scikit-learn;

f). tqdm;

На рисунку 3.2 зображено дерево структури проекту, з ієрархією модулів та програмних папок.

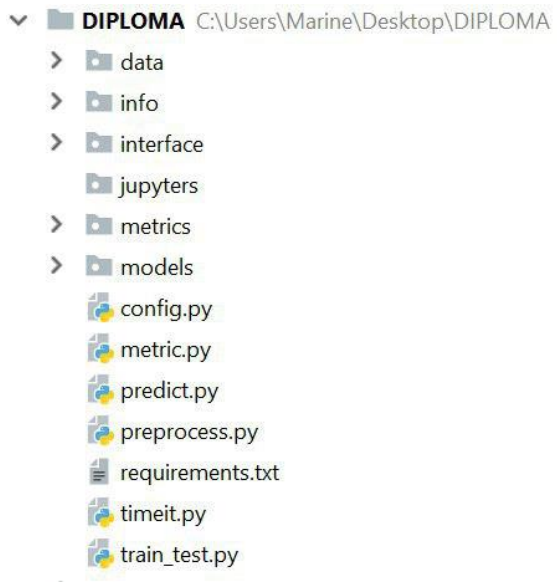


Рисунок 3.2 – Дерево структури проекту

3.3 Побудова моделей демографічних процесів

Після проведення тесту Дікі-Фуллера та Уайта, було виявлено відсутність тренду та класифіковано досліджуваний процес як гетероскедастичний. Тому було проведено порівняння двох основних моделей АРУГ та УАРУГ із застосуванням МНК та РМНК. Методи РМНК дають кращі результати моделювання, а МНК - прогнозування. Про що свідчать ближчий до 1 коефіцієнт детермінації, та ближчий до 2 коефіцієнт Дарвіна-Уотсона у моделей на основі РМНК. Проаналізувавши показники адекватності для обох випадків, робимо висновок, що кращою є модель АРУГ із застосуванням МНК, тому саме вона буде використана при процесі прогнозування в СППР.

3.4 Інструкція по експлуатації програми, аналіз отриманих результатів

3.4.1 Вибір інструментальної платформи для реалізації програми

СППР реалізована мовою програмування python, яка є досить популярною сьогодні. Python нова мова програмування, але не зважаючи на це, вона входить у трійку найпопулярніших мов програмування. До переваг python належать: читабельність, простота та прямота коду, легка інтерпретація, динамічна типізація.

Легка система імпортування, вже вбудовані інструменти та величезна база бібліотек спрощують роботу, і дають можливість сфокусуватись на побудові алгоритму для вирішення конкретних задач.

За допомогою інтерпретатора швидкість проведення експериментів зростає врази. Мова програмування Python надає можливість написання веб додатків, десктоп та мобільних додатків. Тобто це мова широкого призначення.

3.4.2 Інструкція користувача для використання СППР з метою отримання прогнозу

Після запуску СППР на екрані з’явиться вікно, зображене на рисунку 3.3, де у спеціальне поле потрібно ввести назву країни англійською мовою, для якої користувач прагне прогнозувати демографічний стан.

Рисунок 3.3 – Вікно початку роботи програми

Після введення назви країни відкриється вікно, зображене на рисунку 3.4, що відображає результат прогнозу, починаючи з 2013 року і до 2023 року.

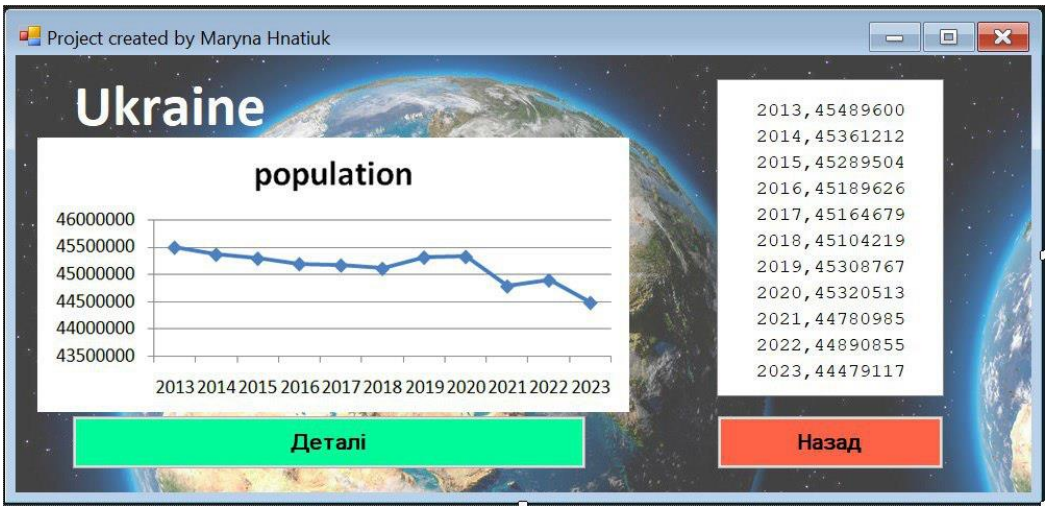


Рисунок 3.4– Результати прогнозування з 2013 по 2023 роки

Для перегляду більш детальної інформації про якість прогнозу користувач може натиснути на кнопку «Деталі», що відкриє вікно, зображене на рисунку 3.5, або ж, повернутися на початкове вікно, для можливості вибору іншої країни світу.

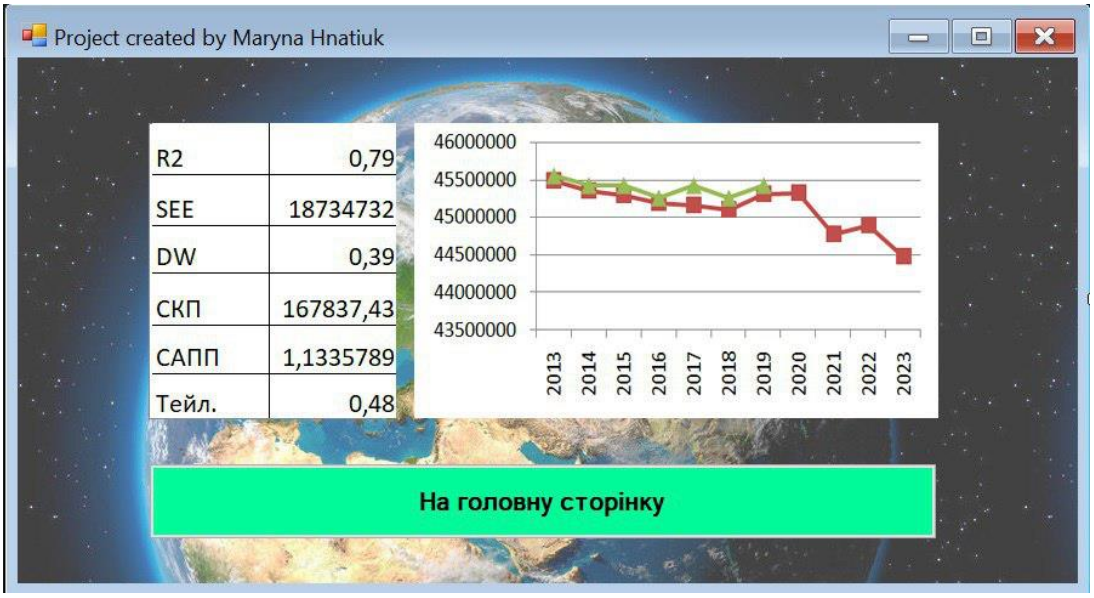


Рисунок 3.5 - Вікно з деталями прогнозування

На даній сторінці можемо спостерігати деякі показники якості прогнозу, а також візуальне порівняння реальних показників 2013 – 2019 років та отриманих результатів прогнозування, що дозволяє наявно оцінити точність роботи СППР.

Як можемо помітити, інтерфейс користувача є максимально простим та зрозумілим. За декілька хвилин, він отримує прогнозовані дані, для обраної ним країни.

Проаналізувавши отримані результати, можемо зробити висновок про те, що характеристики адекватності лежать в межах норми, деякі навіть дають хороший результат, такі як R^2 та Дарбіна-Уотсона. Також слід звернути увагу на візуальне підтвердження точності прогнозу, що зображено на рисунку 3.5 показники незначною мірою відрізняються від реальних статистичних даних.

3.5 Висновки до розділу 3

Після дослідження, збору та обробки статистичних даних, аналізу якості обраних моделей прогнозування та розробки коду, було створено СППР, що дає адекватний прогноз чисельності населення обраної країни світу на N років вперед. СППР була розроблена на основі інформаційного підходу, з використанням гібридного типу архітектури, мовою програмування python.

Було обрано за основу модель АРУГ із використанням МНК, оскільки вона показала кращий результат, ніж УАРУГ за показниками якості моделі. Розроблено максимально простий та зрозумілий інтерфейс користувача, за допомогою якого робота з СППР стала більш приємною.

РОЗДІЛ 4 ФУНКЦІОНАЛЬНО-ВАРТІСНИЙ АНАЛІЗ ПРОГРАМНОГО ПРОДУКТУ

4.1 Вибір методу

Після розробки програмного продукту необхідно провести функціонально-вартісний аналіз основних його характеристик. Це допоможе оцінити проект в цілому, впевнитися у оптимальності вибору середовища розробки та методів, обраних для реалізації поставленої задачі, висвітлити основні сильні сторони програмного продукту.

Розглянемо різні варіації реалізації, для того, щоб обрати найбільш оптимальну з економічної точки зору. Для цього використаємо метод системного дослідження функцій об'єкта з метою пошуку балансу між його собівартістю і корисністю, який має назву функціонально-вартісний аналіз.

Функціонально-вартісний аналіз проводять при розробці та постановці на виробництво нових виробів, підвищенні техніко-економічного рівня продукції. При проектуванні здійснюється пошук оптимальних технічних рішень, встановлення граничних нормативів витрат з виготовлення розроблюваних об'єктів та визначення найбільш ефективних умов їх використання. На етапі виробництва для вдосконалення об’єктів виявляються зайві витрати, визначаються диспропорції між значимістю функцій для споживача і витратами на їх забезпечення, здійснюється пошук резервів для зниження собівартості і підвищення якості виробів.

Крім того, даний метод дозволяє вибрати оптимальний, як з погляду розробника, так і з точки зору покупця варіант розв’язання будь-якої задачі, а також дозволяє оптимізувати витрати й час виконання робіт.

Для реалізації даного методу слід притримуватись наступного алгоритму:

– визначається послідовність функцій, необхідних для виробництва продукту. Спочатку – всі можливі, потім вони розподіляються по двом групам: ті, що впливають на вартість продукту і ті, що не впливають. На цьому ж етапі оптимізується сама послідовність скороченням кроків, що не впливають на цінність і відповідно витрат.

– для кожної функції визначаються повні річні витрати й кількість робочих часів.

– для кожної функції на основі оцінок попереднього пункту визначається кількісна характеристика джерел витрат.

– після того, як для кожної функції будуть визначені їх джерела витрат, проводиться кінцевий розрахунок витрат на виробництво продукту.

4.2 Постановка задачі техніко-економічного аналізу

У даній дипломній роботі було реалізовано систему підтримки прийняття рішень для прогнозування нелінійних нестаціонарних процесів. Розробка велася мовою програмування Python 3.6 у середовищі розробки PyCharm.

СППР відповідає стандартним технічним вимогам:

- Функціональність на ПК, що відповідають вимогам, описаним у третьому розділі даної роботи.
- Однозначний та зрозумілий інтерактивний інтерфейс користувача
- Забезпечення високої швидкості обробки інформації та видачі результатів.
- Передбачення мінімальних витрат на впровадження СППР.

4.2.1 Обґрунтування функцій програмного продукту

Головна функція F_0 – розробка СППР, що аналізує процес за вхідними даними та будує її модель для подальшого прогнозування. Виходячи з конкретної мети, можна виділити наступні основні функції СППР:

F_1 – вибір мови програмування;

F_2 – вибір оптимального середовища розробки;

F_3 – інтерфейс користувача.

Кожна з основних функцій може мати декілька варіантів реалізації.

Функція F_1:

а) мова програмування Java;

б) мова програмування Python.

Функція F_2:

а) PyCharm;

б) IntelliJ IDEA;

Функція F_3:

а) веб додаток;

б) десктопний додаток.

4.2.2 Варіанти реалізації основних функцій

Для опису варіантів реалізації основних функцій було використано морфологічну карту системи, зображену на рисунку 4.1. На її основі було

побудовано позитивно-негативну матрицю варіантів основних функцій, дані якої наведено в таблиці 4.1.

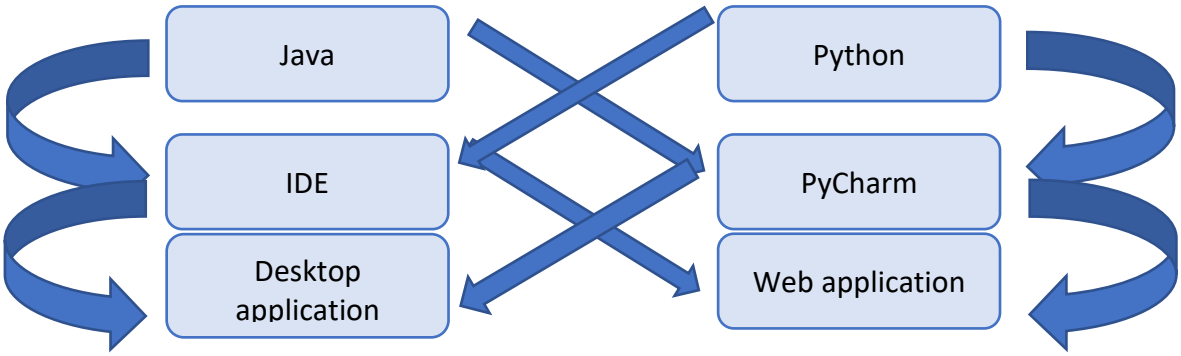


Рисунок 4.1 – Морфологічна карта

Морфологічна карта відображає всі можливі комбінації варіантів реалізації функцій, які складають повну множину варіантів ПП.

Таблиця 4.1 – Позитивно-негативна матриця

Основні функції	**Варіанти реалізації**	**Переваги**	**Недоліки**
F1	А	Структурована архітектура додатку	Складне налаштування додаткових конфігураційних файлів.
	Б	Простий механізм використання бібліотек, що оптимізують роботу програми	Мінімізація часу виконання

F2	А	Зручне середовище програмування з наявністю intellisense	Бібліотеки займають значну кількість пам’яті на жорсткому диску
	Б	Автоматичний збір проекту	Деякий, необхідний для розробки функціонал доступний в небезкоштовній версії
F3	А	Доступність на будь-якому пристрої з доступом до мережі Internet	Для роботи за межами localhost, вимагається використання небезкоштовних серверів
	Б	Зручне та звичне використання на ПК	Обмеження фізичними можливостями ПК

На основі аналізу позитивно-негативної матриці робимо висновок, що при розробці програмного продукту деякі варіанти реалізації функцій варто відкинути, тому, що вони не відповідають поставленим перед програмним продуктом задачам. Ці варіанти відзначені у морфологічній карті.

– Функція F1:

Оскільки в рамках даної задачі налаштування Python не займають багато часу та являються простіші та зрозуміліші у використанні, обираємо варіант Б.

– Функція F2:

При роботі з прогнозуванням дуже зручно використовувати бібліотеки, які значно оптимізують роботу програми. Можливість в один натиск клавіші

автоматично зібрати проект полегшує процес розробки врази. Оскільки їх функціональні можливості дещо схожі, тому обидва варіанти підходять.

– Функція F3:

Оскільки маємо роботу з великою кількістю даних, які необхідно зберігати і оброблювати в процесі моделювання, що займає деякий час, будемо використовувати варіант Б, який не потребує додаткових матеріальних витрат на вищеперераховані пункти, а також не має обмежень по наявності підключення до мережі Internet, що робить його доступність цілодобовою.

Таким чином, будемо розглядати такі варіанти реалізації ПП:

1. F1Б – F2А – F3Б
2. F1Б – F2Б – F3Б

Для оцінювання якості розглянутих функцій обрана система параметрів, описана нижче.

4.3 Обґрунтування системи параметрів ПП

4.3.1 Опис параметрів

На підставі даних про основні функції, що повинен реалізувати програмний продукт, вимог до нього, визначаються основні параметри виробу, що будуть використані для розрахунку коефіцієнта технічного рівня.

Для того, щоб охарактеризувати програмний продукт, будемо використовувати наступні параметри:

– *X1* – простота та зрозумілість мови програмування;

– *X2* – час обробки запиту;

– *X3* – потенційний об’єм програмного коду;

– *X4* – об’єм пам’яті для збереження даних.

X1: Відображає логічну навантаженість сенсу функцій та мінімізує час на розуміння їхнього призначення розробниками залежно від обраної мови програмування.

X2: Відображає час, який витрачається на дії.

*X3:*Показує розмір програмного коду, який необхідно створити розробнику.

X4: Відображає об’єм пам’яті в оперативній пам’яті персонального комп’ютера, необхідний для збереження та обробки даних під час виконання програми.

4.3.2 Кількісна оцінка параметрів

Гірші, середні і кращі значення параметрів вибираються на основі вимог замовника й умов, що характеризують експлуатацію СППР як показано у таблиці 4.2.

Таблиця 4.2 – Основні параметри ПП

Назва Параметра	**Умовні позначення**	**Одиниці виміру**	**Значення параметра**		
			гірші	**середні**	**кращі**
Простота та зрозумілість мови програмування	X1	%	0	50	100
Час обробки запитів	X2	мс	1500	500	50
Потенційний об’єм прогр. Коду	X3	кількість строк коду	500	200	100

Об’єм пам’яті для збереження даних.	X4	Кб	1000	500	300

За даними таблиці 4.2 будуються графічні характеристики параметрів – рисунки 4.2 – 4.5

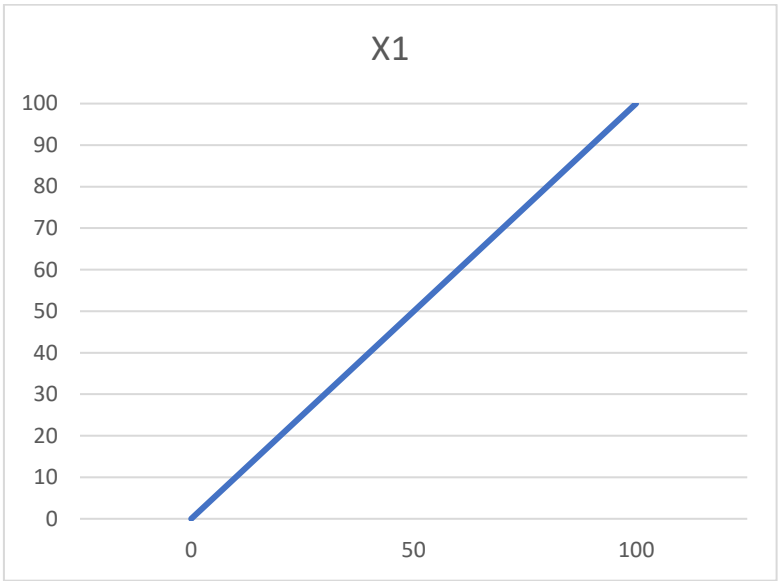


Рисунок 4.2 – X1 - Простота та зрозумілість мови програмування

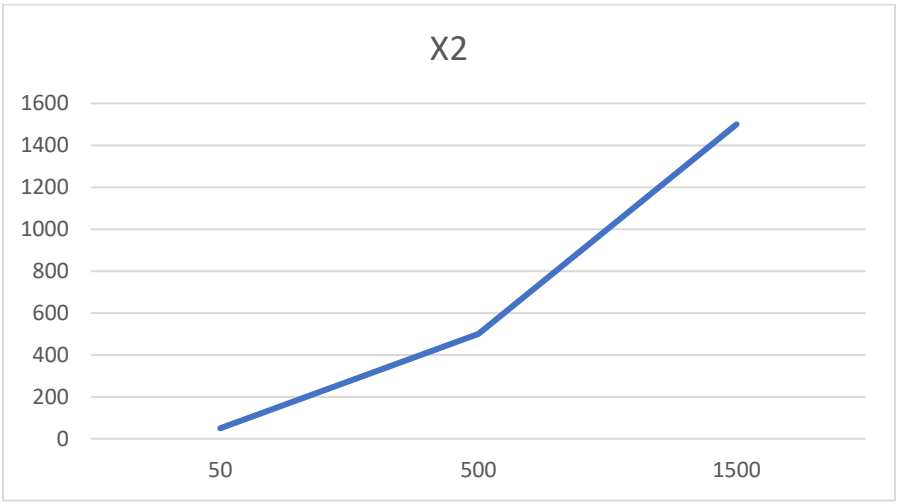


Рисунок 4.3 – X2 - Час обробки запитів

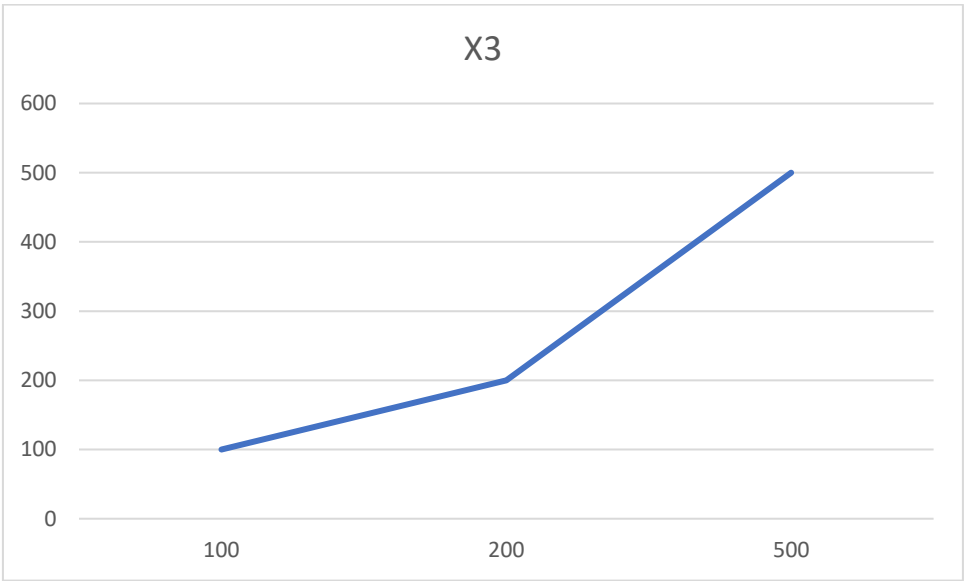


Рисунок 4.4 – X3 - Потенційний об’єм програмного коду

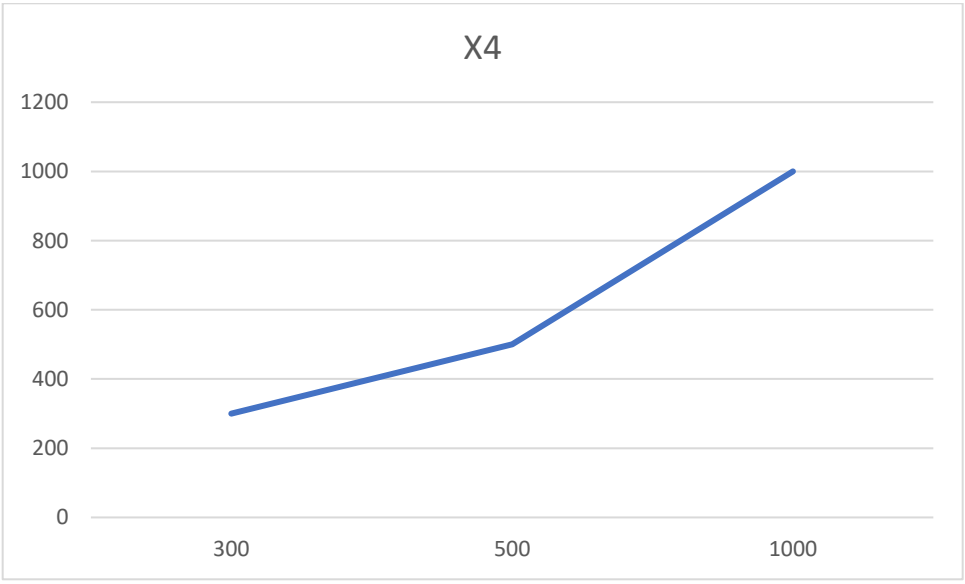


Рисунок 4.5 – X4 - Об’єм пам’яті для збереження даних.

4.3.3 Аналіз експертного оцінювання параметрів

Після детального обговорення й аналізу кожний експерт оцінює ступінь важливості кожного параметру для конкретно поставленої цілі – розробка програмного продукту, який дає найбільш точні результати при знаходженні параметрів моделей адаптивного прогнозування і обчислення прогнозних значень.

Значимість кожного параметра визначається методом попарного порівняння. Оцінку проводить експертна комісія із 7 людей. Визначення коефіцієнтів значимості передбачає:

– визначення рівня значимості параметра шляхом присвоєння різних рангів;

– перевірку придатності експертних оцінок для подальшого використання;

– визначення оцінки попарного пріоритету параметрів;

– обробку результатів та визначення коефіцієнту значимості.

Результати експертного ранжування наведені у таблиці 4.3.

Таблиця 4.3 – Результати ранжування параметрів

Позначення параметра	Назва параметра	Одиниці виміру	Ранг параметра за оцінкою експерта							Сума рангів R_i	Відхилення Δ_i	Δ_i^2
			1	2	3	4	5	6	7			
X1	Простота та зрозумілість мови програмування	%	1	3	4	1	4	2	2	17	-0.5	0.25

X2	Час обробки запитів	Мс	2	2	1	2	1	3	3	14	-3.5	12.25
X3	Потенційний об'єм програмного коду	кількість строк коду	4	1	2	3	2	1	1	14	-3.5	12.25
X4	Об'єм пам'яті для збереження даних.	Кб	3	4	3	4	3	4	4	25	7.5	56.25
	Разом		10	10	10	10	10	10	10	70	0	429,656

Для перевірки степені достовірності експертних оцінок, визначимо наступні параметри:

а) сума рангів кожного з параметрів і загальна сума рангів:

$$R_i = \sum_{j=1}^{N} r_{ij} R_{ij} = \frac{Nn(n+1)}{2} = 70, \quad (4.1)$$

де N – число експертів;

n – кількість параметрів;

б) середня сума рангів:

$$T = \frac{1}{n} R_{ij} = 17.5 \quad (4.2)$$

в) відхилення суми рангів кожного параметра від середньої суми рангів:

$$\Delta_i = R_i - T \tag{4.3}$$

Сума відхилень по всім параметрам повинна дорівнювати 0;

г)загальна сума квадратів відхилення:

$$S = \sum_{i=1}^{N} \Delta_i^2 = 429{,}656 \tag{4.4}$$

Порахуємо коефіцієнт узгодженості:

$$W = \frac{12S}{N^2(n^3-n)} = \frac{12 \cdot 429{,}656}{7^2(4^3-4)} = 1{,}75 > W_{\text{норм}} = 0{,}67 \tag{4.5}$$

Ранжування можна вважати достовірним, тому що знайдений коефіцієнт узгодженості перевищує нормативний, котрий дорівнює 0,67.

Скориставшись результатами ранжирування, проведемо попарне порівняння всіх параметрів і результати занесемо у таблицю 4.4.

Таблиця 4.4 – Попарне порівняння параметрів

Параметри	Експерти							Кінцева оцінка	Числове значення
	1	2	3	4	5	6	7		
X1 і X2	<	>	>	<	>	<	<	<	0,5
X1 і X3	<	>	>	<	>	>	>	>	1,5
X2 і X3	<	>	<	<	<	>	>	<	0,5
X1 і X4	<	<	>	<	>	<	<	<	0,5
X2 і X4	<	<	<	<	<	<	<	<	0,5
X3 і X4	>	<	<	<	<	<	<	<	0,5

Числове значення, що визначає ступінь переваги *i*–го параметра над *j*–тим, a_{ij} визначається по формулі:

$$a_{ij} = \begin{cases} 1.5, \text{при Xi} > \text{Xj} \\ 1.0, \text{при Xi} = \text{Xj} \\ 0.5, \text{при Xi} < \text{Xj} \end{cases} \quad (4.6)$$

З отриманих числових оцінок переваги складемо матрицю А=‖ a_{ij} ‖.Для кожного параметра зробимо розрахунок вагомості $K_{вi}$ за наступними формулами:

$$K_{\text{вi}} = \frac{b_i}{\sum_{i=1}^{n} b_i}, \quad (4.7)$$

де $b_i = \sum_{i=1}^{N} a_{ij}$.

Відносні оцінки розраховуються декілька разів доти, поки наступні значення не будуть незначно відрізнятися від попередніх (менше 2%).

На другому і наступних кроках відносні оцінки розраховуються за наступними формулами:

$$K_{\text{вi}} = \frac{b_i'}{\sum_{i=1}^{n} b_i'}, \text{де } b_i' = \sum_{i=1}^{N} a_{ij} b_j. \quad (4.8)$$

Як видно з таблиці 4.5, різниця значень коефіцієнтів вагомості не перевищує 2%, тому більшої кількості ітерацій не потрібно.

Таблиця 4.5 – Розрахунок вагомості параметрів

Параметри x_i	Параметри x_j				Перша ітер.		Друга ітер.		Третя ітер.	
	X1	X2	X3	X4	b_i	$K_{\text{ві}}$	b_i^1	$K_{\text{ві}}^1$	b_i^2	$K_{\text{ві}}^2$
X1	1,0	0,5	1,5	0,5	3,5	0.2187	12,25	0.1828	44.875	0.2111
X2	1,5	1,0	0,5	0,5	3,5	0.2187	12,25	0.1828	44.875	0.2111
X3	0,5	1,5	1,0	0,5	3,5	0.2187	12,25	0.1828	44.875	0.2111
X4	1,5	1,5	1,5	1.0	5,5	0.3437	30,25	0.4514	77.875	0.3664
Всього:					16	1	67	1	212.5	1

Визначаємо рівень якості кожного варіанту виконання основних функцій окремо. Коефіцієнт технічного рівня для кожного варіанта реалізації СППР розраховується так (таблиця 4.6):

$$K_K(j) = \sum_{i=1}^{n} K_{\text{в}i,j} B_{i,j}, \tag{4.9}$$

де n – кількість параметрів;

$K_{\text{в}i}$– коефіцієнт вагомості i–го параметра;

B_i – оцінка i–го параметра в балах.

Таблиця 4.6 – Розрахунок показників рівня якості варіантів реалізації основних функцій СППР

Основні функції	Варіант реалізації функції	Параметри x_i	Абсолютне значення параметра	Бальна оцінка параметра	Коефіцієнт вагомості параметра	Коефіцієнт рівня якості

F1	б)	Xl	80	5	0.2111	1,055
F3	б)	X4	200	4	0.3664	1,4656
F2	б)	X2	100	3	0.2111	0,6333
		X3	150	2	0.2111	0,4222
	а)	X2	1000	3	0.2111	0,6333
		X3	450	3	0.2111	0,6333

За даними з таблиці 4.6 за формулою

$$K_K = K_{\text{ТУ}}[F_{1k}] + K_{\text{ТУ}}[F_{2k}] + \ldots + K_{\text{ТУ}}[F_{zk}], \quad (4.10)$$

визначаємо рівень якості кожного з варіантів:

K_{K1} = 1,055+0,6333+0,6333+1,4656= 3,783

K_{K2} = 1,055+0,4222+0,6333+1,4656= 3,576

Як видно з розрахунків, кращим є перший варіант, для якого коефіцієнт технічного рівня має найбільше значення.

4.4 Економічний аналіз варіантів розробки СППР

Для визначення вартості розробки СППР спочатку проведемо розрахунок трудомісткості.

Всі варіанти включають в себе два окремих завдання:

1. Розробка системи підтримки прийняття рішень;

2. Написання алгоритму обробки даних та побудови моделі

Завдання 1 за ступенем новизни відноситься до групи А, завдання 2 – до групи Б. За складністю алгоритми, які використовуються в завданні 1 належать до групи 1; а в завданні 2 – до групи 3. Для реалізації завдання 1 використовується довідкова інформація, а завдання 2 використовує інформацію у вигляді даних.

Проведемо розрахунок норм часу на розробку та програмування для кожного з завдань.

Проведемо розрахунок норм часу на розробку та програмування для кожного з завдань. Загальна трудомісткість обчислюється як

$$T_O = T_P \cdot K_П \cdot K_{СК} \cdot K_M \cdot K_{СТ} \cdot K_{СТ.М}, \quad (4.11)$$

де T_P – трудомісткість розробки ПП;

$K_П$ – поправочний коефіцієнт;

$K_{СК}$ – коефіцієнт на складність вхідної інформації;

K_M – коефіцієнт рівня мови програмування;

$K_{СТ}$ – коефіцієнт використання стандартних модулів і прикладних програм;

$K_{СТ.М}$ – коефіцієнт стандартного математичного забезпечення

Для першого завдання, виходячи із норм часу для завдань розрахункового характеру степеню новизни А та групи складності алгоритму 1, трудомісткість дорівнює: T_P =90 людино-днів. Поправочний коефіцієнт, який враховує вид нормативно-довідкової інформації для першого завдання: $K_П$ = 1.7. Поправочний коефіцієнт, який враховує складність контролю вхідної та вихідної інформації для завдань рівний 1: $K_{СК}$ = 1. Оскільки при розробці першого завдання використовуються стандартні модулі, врахуємо це за допомогою коефіцієнта $K_{СТ}$

= 0.8. Тоді, за формулою (4.11), загальна трудомісткість програмування першого завдання дорівнює:

$$T_1 = 90 \cdot 1.7 \cdot 0.8 = 122.4 \text{ людино-днів.}$$

Проведемо аналогічні розрахунки для подальших завдань.

Для другого завдання (використовується алгоритм третьої групи складності, степінь новизни Б), тобто $T_Р$ =27 людино-днів, $К_П$ =0.9, $К_{СК}$ = 1, $К_{СТ}$ =0.8:

$$T_2 = 27 * 0.9 * 0.8 = 19.44 \text{ людино-днів.}$$

Складаємо трудомісткість відповідних завдань для кожного з обраних варіантів реалізації програми, щоб отримати їх трудомісткість:

$$T_I = (122.4 + 19.44) * 15 = 2127{,}6 \text{ людино-годин;}$$
$$T_{II} = (122.4 + 19.44 + 9) * 15 = 2262{,}6 \text{ людино-годин;}$$

Найбільш високу трудомісткість має варіант II.

В розробці беруть участь два програмісти з окладом 14000 грн., один спеціаліст по цифровій обробці сигналів з окладом 22000грн. Визначимо зарплату за годину за формулою:

$$СЧ = \frac{М}{T_m \cdot t} \text{ грн.,} \qquad (4.12)$$

де М – місячний оклад працівників;

T_m – кількість робочих днів тиждень;

t – кількість робочих годин в день.

$$СЧ = \frac{14000 + 14000 + 22000}{3 \cdot 21 \cdot 15} = 52{,}91 \text{ грн.}$$

Тоді, розрахуємо заробітну плату за формулою

$$СЗП = С_ч \cdot Т_i \cdot КД, \quad (4.13)$$

де $С_ч$ – величина погодинної оплати праці програміста;

$Т_i$ – трудомісткість відповідного завдання;

$К_Д$ – норматив, який враховує додаткову заробітну плату.

Зарплата розробників за варіантами становить:

I. $С_{ЗП} = 52{,}91 \cdot 2127{,}6 \cdot 1.2 = 135085{,}58$ грн.

II. $С_{ЗП} = 52{,}91 \cdot 2262{,}6 \cdot 1.2 = 143657{,}00$ грн.

Відрахування на єдиний соціальний внесок в залежності від групи професійного ризику (II клас) становить 22%:

I. $С_{ВІД} = С_{ЗП} \cdot 0.3677 = 135085{,}58 \cdot 0.22 = 29718{,}83$ грн.

II. $С_{ВІД} = С_{ЗП} \cdot 0.3677 = 143657{,}00 \cdot 0.22 = 31604{,}54$ грн.

Тепер визначимо витрати на оплату однієї машино-години. ($С_М$)

Так як одна ЕОМ обслуговує одного програміста з окладом 14000 грн., з коефіцієнтом зайнятості 0,2 то для однієї машини отримаємо:

$$С_Г = 12 \cdot М \cdot К_З = 12 \cdot 12000 \cdot 0{,}2 = 28800 \text{ грн.}$$

З урахуванням додаткової заробітної плати:

$$С_{ЗП} = С_Г \cdot (1 + К_З) = 28800 \cdot (1 + 0.2) = 34560 \text{ грн.}$$

Відрахування на єдиний соціальний внесок:

$$С_{ВІД} = С_{ЗП} \cdot 0.3677 = 34560 \cdot 0,22 = 7603,20 \text{ грн.}$$

Амортизаційні відрахування розраховуємо при амортизації 25% та вартості ЕОМ – 8000 грн.

$$С_А = К_{ТМ} \cdot К_А \cdot Ц_{ПР} = 1.15 \cdot 0.25 \cdot 8000 = 2300 \text{ грн.},$$

де $К_{ТМ}$– коефіцієнт, який враховує витрати на транспортування та монтаж приладу у користувача;

$К_А$– річна норма амортизації;

$Ц_{ПР}$– договірна ціна приладу.

Витрати на ремонт та профілактику розраховуємо як:

$$С_Р = К_{ТМ} \cdot Ц_{ПР} \cdot К_Р = 1.15 \cdot 8000 \cdot 0.05 = 460 \text{ грн.},$$

де $К_Р$– відсоток витрат на поточні ремонти.

Ефективний годинний фонд часу ПК за рік розраховуємо за формулою:

$$Т_{ЕФ} = (Д_К – Д_В – Д_С – Д_Р) \cdot t_З \cdot К_В \qquad (4.14)$$

$$Т_{ЕФ} = (365 – 104 – 8 – 16) \cdot 8 \cdot 0.9 = 1706.4 \text{ годин},$$

де $Д_К$ – календарна кількість днів у році;

$Д_В$, $Д_С$ – відповідно кількість вихідних та святкових днів;

$Д_Р$ – кількість днів планових ремонтів устаткування;

t –кількість робочих годин в день;

$К_В$– коефіцієнт використання приладу у часі протягом зміни.

Витрати на оплату електроенергії розраховуємо за формулою:

$$С_{ЕЛ} = Т_{ЕФ} \cdot N_С \cdot К_З \cdot Ц_{ЕН} = 1706,4 \cdot 0,156 \cdot 0,9733 \cdot 2,7515 = 712,88 \text{ грн.},$$

де $N_С$ – середньо-споживча потужність приладу;

$К_З$– коефіцієнтом зайнятості приладу;

$Ц_{ЕН}$ – тариф за 1 КВт-годин електроенергії.

Накладні витрати розраховуємо за формулою:

$$С_Н = Ц_{ПР} \cdot 0.67 = 8000 \cdot 0,67 = 5360 \text{ грн.}$$

Тоді, річні експлуатаційні витрати будуть:

$$С_{ЕКС} = С_{ЗП} + С_{ВІД} + С_А + С_Р + С_{ЕЛ} + С_Н \qquad (4.15)$$

$$С_{ЕКС} = 34560 + 7603,20 + 2300 + 460 + 712,88 + 5360 = 50996,08 \text{ грн.}$$

Собівартість однієї машино-години ЕОМ дорівнюватиме:

$$С_{М\text{-}Г} = С_{ЕКС} / Т_{ЕФ} = 50996,08/1706,4 = 29,885 \text{ грн/час.}$$

Оскільки в даному випадку всі роботи, які пов‘язані з розробкою програмного продукту ведуться на ЕОМ, витрати на оплату машинного часу, в залежності від обраного варіанта реалізації, складає:

$$C_М = C_{М\text{-}Г} \cdot Т \quad (4.15)$$

I. $C_М = 29{,}885 \cdot 2127{,}6 = 63583{,}326$ грн.;

II. $C_М = 29{,}885 \cdot 2262{,}6 = 67617{,}801$ грн.;

Накладні витрати складають 67% від заробітної плати:

$$C_Н = C_{ЗП} \cdot 0{,}67 \quad (4.16)$$

I. $C_Н = 135085{,}58 \cdot 0{,}67 = 90507{,}34$ грн.;

II. $C_Н = 143657{,}00 \cdot 0{,}67 = 96250{,}19$ грн.;

Отже, вартість розробки ПП за варіантами становить:

$$C_{ПП} = C_{ЗП} + C_{ВІД} + C_М + C_Н \quad (4.17)$$

I. $C_{ПП} = 135085{,}58 + 49670{,}97 + 63583{,}326 + 90507{,}34 = 338847{,}216$ грн.;

II. $C_{ПП} = 143657{,}00 + 52822{,}68 + 67617{,}801 + 96250{,}19 = 360347{,}671$ грн.;

4.5 Вибір кращого варіанта СППР за техніко-економічним рівнем

Розрахуємо коефіцієнт техніко-економічного рівня за формулою:

$$К_{ТЕРj} = К_{Кj} / С_{Фj}, \quad (4.18)$$

$К_{ТЕР1}$ = 3,783 / 344325,79 = $10,99 \cdot 10^{-6}$;

$К_{ТЕР2}$ =3,576 / 366173,97= $9,77 \cdot 10^{-6}$;

Як бачимо, найбільш ефективним є перший варіант реалізації програми з коефіцієнтом техніко-економічного рівня $К_{ТЕР1}= 10,99 \cdot 10^{-6}$.

4.6 Висновки до розділу 4

В даній розрахунковій роботі проведено повний функціонально-вартісний аналіз СППР, який було розроблено в рамках дипломної роботи. Процес аналізу можна умовно розділити на дві частини.

В першій з них проведено дослідження СППР з технічної точки зору: було визначено основні функції СППР та сформовано множину варіантів їх реалізації; на основі обчислених значень параметрів, а також експертних оцінок їх важливості було обчислено коефіцієнт технічного рівня, який і дав змогу визначити оптимальну з технічної точки зору альтернативу реалізації функцій СППР.

Другу частину ФВА присвячено вибору із альтернативних варіантів реалізації найбільш економічно обґрунтованого. Порівняння запропонованих варіантів реалізації в рамках даної частини виконувалось за коефіцієнтом ефективності, для обчислення якого були обчислені такі допоміжні параметри, як трудомісткість, витрати на заробітну плату, накладні витрати.

Після виконання функціонально-вартісного аналізу програмного комплексу що розроблюється, можна зробити висновок, що з альтернатив, що

залишились після першого відбору двох варіантів виконання програмного комплексу оптимальним є перший варіант реалізації програмного продукту. У нього виявився найкращий показник техніко-економічного рівня якості $К_{ТЕР}$ $10{,}99 \cdot 10^{-6}$.

Цей варіант реалізації програмного продукту має такі параметри:

- мова програмування – Python;
- використання PyCharm;
- реалізація Desktop application.

Даний варіант виконання програмного комплексу дає користувачу зрозумілий та ефективний функціонал, швидкодію і робить простішим виконання завдання.

ВИСНОВКИ

Кожного дня наша планета поповнюється новими жителями і, на жаль, також несе втрати, відбуваються міграції з одного місця в інше, і навпаки, саме ці динамічні процеси творять наше життя. Вміння спрогнозувати їх поведінку дає величезні можливості для побудови стратегій ведення внутрішньої та зовнішньої політики держав, таким чином покращити рівень життя, та піднятись у рейтингу на світовій арені.

У даній роботі було проаналізовано демографічні процеси нашої країни та ще однієї провідної країни Європи – Німеччини. Досліджено ряд моделей прогнозування нелінійних нестаціонарних процесів та показників їх якості. Висвітлено переваги та недоліки вже існуючих популярних пакетних рішень створення прогнозу.

Було сформовано методику побудови моделей інтегрованих та гетероскедастичних процесів, обрано найкращу модель для прогнозування демографічних процесів та найголовніше – реалізовано СППР, що має простий, зрозумілий, дружній інтерфейс користувача, дає можливість спрогнозувати чисельність населення обраної країни світу за короткий період часу, є гнучкою у своєму використанні, написана на новій, популярній мові програмування python, підходить для різних видів ОС.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Бідюк П.І., Половцев О.В. Аналіз та моделювання економічних процесів перехідного періоду / К: ПЛАБ-75, 1999. 230 с.

2. Бідюк П.І. Системний підхід до побудови математичних моделей на основі часових рядів / Системні дослідження та інрормаційні технології, №3, 2002. l 14-131 с.

3. Бідюк П.І. Часові ряди: моделювання та прогнозування / Київ: ЕКМО, 2004.144 с.

4. Шекера О.Г. Демографічна ситуація у світі та в Україні / Науково-практичний журнал «Здоров'я суспільства», 2014.

5. Власенко Н.С., Макарова О.В., Пирожков С.І., та інші. Комплексний демографічний прогноз України на період до 2050 р. / за ред. член-кореспондент НАНУ, д.е.н., проф. Е.М. Лібанової. К.: Український центр соціальних реформ, 2006. 138 с.

6. Цвігун І.А. Демографічна безпека України та напрями її регулювання: монографія / Кам’янець-Подільський: Видавець ПП Зволейко Д.Г., 2013. 400 с.

7. Пальян З. О. Навчальний посібник Демографічна статистика / Київський Національний Економічний Університет України, 2003 р. 167 с.

8. Бідюк П.І., Коршевнюк Л.О., Проектування комп'ютерних інформаційних систем підтримки прийняття рішень: Навчальний посібник / Київ: ННК «Інститут прикладного системного аналізу» Національний технічний університет України «Київський політехнічний інститут», 2010. 340 с.

9. Пашин В.П. Функционально-стоимостный анализ конструкторско-технологических решений. - К.: РДЭНТП «Знание» УССР, 1989. - 22с.

10. Бідюк П.І., Коршевнюк Л.О. Проектування комп’ютерних інформаційних систем підтримки прийняття рішень: Навчальний посібник. Київ: ННК «ІПСА» НУТУ «КПІ», 2010. 340 с.

11. П. І. Бідюк, В. Д. Романенко, О. Л.Тимощук Аналіз часових рядів (навчальний посібник). К.: Політехніка, 2010. 317 с.

12. Раздел журнала: Методы обработки информации. УДК 62-50П.И. Бидюк, А.Н. Терентьев, Л.А. Коршевнюк байесовская сеть - инструмент интеллектуального анализа данных.

13. Бидюк П.И., Терентьев А.Н., Гасанов А.С. Построение и методы обучения, Кибернетика и системный анализ, 2005, № 4. с. 133 – 147.

14. Challis, R. E., and Kitney, R. I. (November 1991). "Biomedical signal processing (in four parts). Part 1 Time-domain methods." Medical & Biological Engineering & Computing, 28, 509-524.

15. Лукашин Ю.П. Адаптивные методы долгосрочного прогнозирования. М.: Финансы и статистика, 2003. 414 с.

16. Зельнер А. Методы в эконометрии. Москва: Статистика, 1980. 438 с.

17. Chatfield C. Time series forecasting. London: Chapman & Hall, 2000. 267 p.

18. Згуровский М.З., Подладчиков В.Н. Аналитические методы калмановской фильтрации. Київ: Наукова думка, 1995. 285 с.

19. Fryzlewicz, P., Van Bellegem, S. & von Sachs, R 2003. Forecasting non-stationary time series by wavelet process modelling. Annals of the Institute of Statistical Mathematics, 55, 737-764. Та Priestley, M.B. 1983. Spectral Analysis and Time Series. Academic Press.

ДОДАТОК А ІЛЮСТРАТИВНИЙ МАТЕРІАЛ

Система підтримки прийняття рішень для прогнозування демографічних процесів

Гнатюк Марина Степанівна, гр. КА-55
Керівник: проф. д.т.н. Бідюк Петро Іванович

1

Об'єкт дослідження:

Демографічні процеси, чисельність та тенденція розвитку популяції населення різних країн світу

2

Предмет дослідження:

Нелінійні нестаціонарні моделі для прогнозування демографічних процесів

3

Мета дослідження:

Побудова СППР для забезпечення прогнозу демографічних процесів

4

Актуальність:

- Демографічні процеси мають прямий вплив на економічний стан країн світу.
- Економіка і Демографія визначають загальний рейтинг держав на міжнародній арені
- Створення СППР, що є легкою у застосуванні, і дає можливість отримати досить точний прогноз є хорошим інструментом для аналізу напрямку ведення економічної політики

5

Постановка задачі

Розробка власної СППР

Створення та вдосконалення моделі демографічних процесів

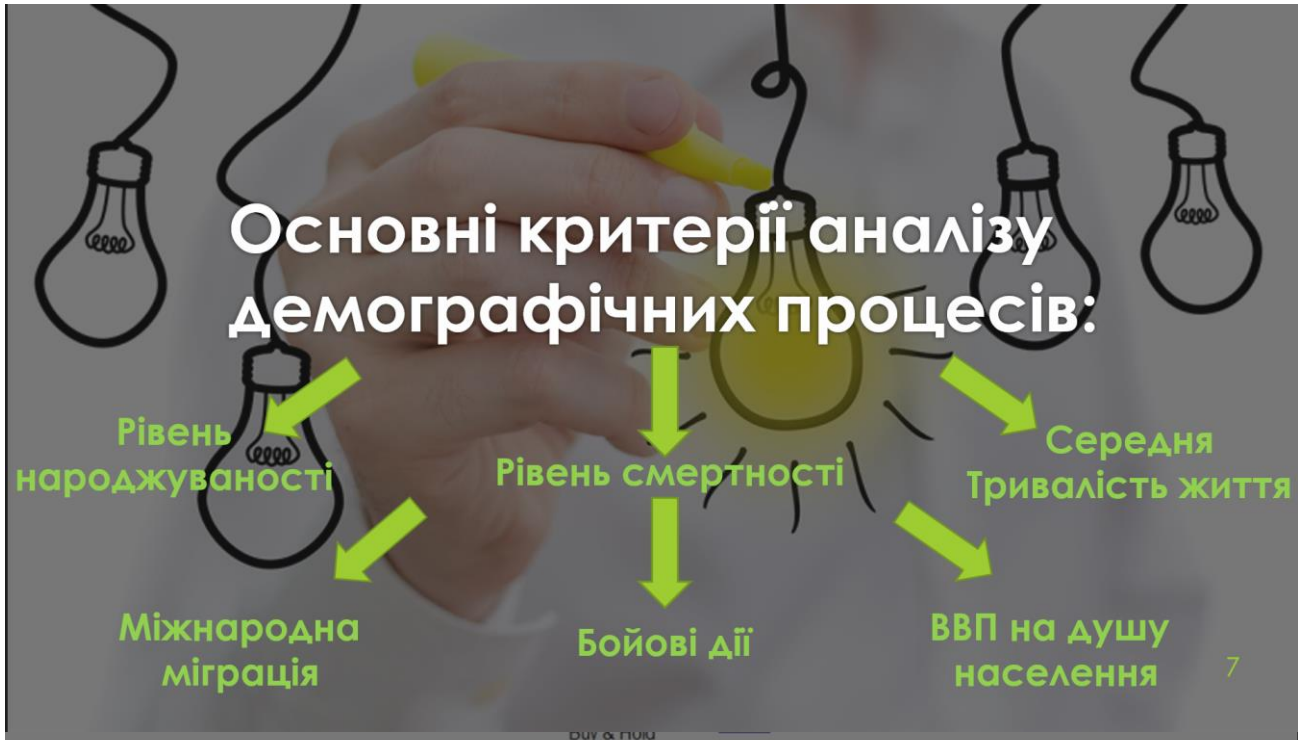
Аналіз популярних системи для побудови моделей стаціонарних і нестаціонарних процесів

Дослідження деяких моделей для нелінійних нестаціонарних процесів

Збір та аналіз статистичних даних, створення датасету

Дослідження демографічних процесів та виділення їх основних характеристик

6

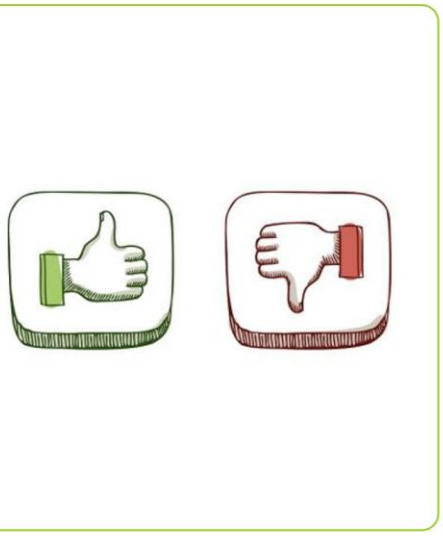

Основні критерії аналізу демографічних процесів:
Рівень народжуваності
Рівень смертності
Середня Тривалість життя
Міжнародна міграція
Бойові дії
ВВП на душу населення
7


Деякі розглянуті моделі:
Авторегресія з трендовою складовою:
$y(k) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + \sum_{j=1}^{m} b_j k^j + \varepsilon(k)$
Поліноміальна модель
$y(k) = a_0 + a_1 k + a_2 k^2 + \cdots + a_m k^m + \varepsilon(k)$
Модель з кубічним сплайном
$S_i(x) = a_i + b_i(x - x_i) + \frac{c_i}{2}(x - x_i)^2 + \frac{d_i}{6}(x - x_i)^3$
Експоненційна модель
$y(k) = ap^k + \varepsilon(k)$
Модель з циклічним компонентом
$y(k) = a_0 + \sum_{i=1}^{p} a_i sin(\omega_0 i) + \sum_{j=1}^{q} b_j cos(\omega_0 i) + \varepsilon(k)$
8

Переваги та недоліки пакету Eviews:

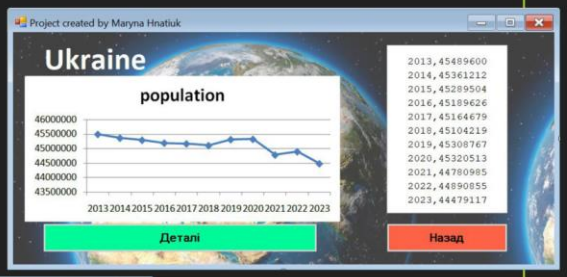
1. проведення всебічного аналізу даних;
2. проведення різних типів прогнозів;
3. прогноз розвитку економічних процесів проблемної області;

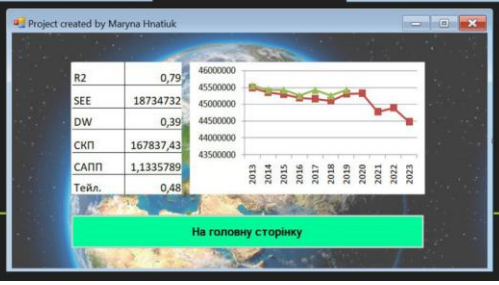
1. нагромадженість не завжди потрібними функціями;
2. складний і незрозумілий інтерфейс;
3. існування нових інструментів, що дозволять аналізувати дані врази швидше;
4. відсутність інтеграції з іншими операційними системи;



9

Процес роботи СППР:





10

Аналіз результатів:

*Отриманий результат дуже схожий із тим, що видав Eviews

Ukraine	Germany	Poland	China
2014,45361212	2013,80645605	2013,38540196	2013,1357380000
2015,45289504	2014,81236284	2014,38481981	2014,1364026083
2016,45189626	2015,82182587	2015,38445089	2015,1371026083
2017,45164679	2016,82549909	2016,38438260	2016,1379708390
2018,45104219	2017,82536425	2017,38439866	2017,1386023312
2019,45308767	2018,82703950	2018,38440598	2018,1395160614
2020,45320513	2019,83026888	2019,38491296	2019,1392890170
2021,44780985	2020,83430552	2020,38569243	2020,1393963604
2022,44890855	2021,83975187	2021,38605472	2021,1401141145
2023,44479117	2022,83744998	2022,38632619	2022,1405378811
	2023,83883404	2023,38657374	2023,1410578811

Висновки:

- Досліджено демографічні процеси та виділено їх основні характеристики
- Зібрано статистичні дані
- Досліджено деякі моделі для нелінійних нестаціонарних процесів
- Виділено переваги і недоліки сучасних систем для моделювання процесів
- Розроблено власну СППР, що дає можливість спрогнозувати чисельність населення на 5-10 років вперед

12

Дякую за увагу!

ДОДАТОК Б ЛІСТИНГ ПРОГРАМИ

Config.py

```
import os

from os.path import join as path_join

APP_DIR = str(os.path.dirname(os.path.abspath(file)))
DATA_DIR = path_join(APP_DIR, 'data/')
METRIC_DIR = path_join(APP_DIR, 'metrics/')
MODELS_DIR = path_join(APP_DIR, 'models/')
```

Metric.py

```
import numpy as np

from typing import Sequence
from sklearn.metrics import mean_squared_log_error

def rmsle(ytrue: Sequence, ypred: Sequence):
    """ Count root mean squared log error """
    return np.sqrt(mean_squared_log_error(ytrue, ypred))
```

Predict.py

```
import pickle
import pandas as pd
import numpy as np

from os.path import join as path_join
from tqdm import tqdm

from config import DATA_DIR, MODELS_DIR
from preprocess import feature_extraction, melt_data


MODEL = 'lgb.pkl'

COUNTRY = 'Ukraine'
FOR_THE_NEXT = 10  # years


if name == 'main':
    model = pickle.load(open(path_join(MODELS_DIR, MODEL), 'rb'))
```

```
    population_data = pd.read_excel(path_join(DATA_DIR, 'Country-Population.xls'))
    melted = melt_data(data=population_data.drop(
        ['Country Code', 'Indicator Name', 'Indicator Code'], axis=1,
    ))
    melted = melted.loc[melted['Country Name'] == COUNTRY]
    print(melted.columns)
    features = feature_extraction(melted, id_column='Country Name')

    for i in tqdm(range(1, FOR_THE_NEXT+1, 1), total=FOR_THE_NEXT, desc='Predict'):
        pred_population = np.expm1(model.predict(features.iloc[-1:].drop(labels=['Country Name',
'year', 'population'], axis=1)))
        new_sample = {
            'Country Name'     : COUNTRY,
            'year'             : features.iloc[-1]['year'] + 1,
            'population'       : int(pred_population[0]),
        }

        melted = melted[['Country Name', 'year', 'population']].append(pd.Series(new_sample),
ignore_index=True)
        features = feature_extraction(melted, id_column='Country Name')
    features[['year', 'population']].to_csv(path_join(DATA_DIR, f'{COUNTRY}_predictions.csv'),
index=False)
```

Preprocess.py

```
import pandas as pd

from os.path import join as path_join
from typing import Tuple

from config import DATA_DIR


def melt_data(data: pd.DataFrame) -> pd.DataFrame:
    """ Melt data """
    df = data.copy(deep=True)
    df = df.melt(id_vars='Country Name', var_name='year', value_name='population')
    df['year'] = df['year'].str.extract('(\d+)', expand=False).astype(int)
    df = df.sort_values(by='Country Name').fillna(method='ffill').sort_values(by=['year', 'Country
Name'])
    df['population'] = df.population.astype(int)
    melt = df.sort_values(['year', 'Country Name'])
    return melt
```

```
def feature_extraction(data: pd.DataFrame, id_column: str = 'id') -> pd.DataFrame:
    """ Extract features from the data """
    df = data.copy(deep=True)
    df['last_year_pop'] = df.groupby([id_column])['population'].shift()
    df['last_year_diff'] = df.groupby([id_column])['last_year_pop'].diff()
    df['last-1_year_pop'] = df.groupby([id_column])['population'].shift(2)
    df['last-1_year_diff'] = df.groupby([id_column])['last-1_year_pop'].diff()
    df['last-2_year_pop'] = df.groupby([id_column])['population'].shift(3)
    df['last-2_year_diff'] = df.groupby([id_column])['last-2_year_pop'].diff()
    df['last-3_year_pop'] = df.groupby([id_column])['population'].shift(4)
    df['last-3_year_diff'] = df.groupby([id_column])['last-3_year_pop'].diff()
    df['last-4_year_pop'] = df.groupby([id_column])['population'].shift(5)
    df['last-4_year_diff'] = df.groupby([id_column])['last-4_year_pop'].diff()
    df = df.dropna()
    return df


def get_train_test(data: pd.DataFrame, test_window: int = 5) -> Tuple[pd.DataFrame,
pd.DataFrame]:
    """ Create is test column """
    max_ = data.year.max()

    train = data.loc[data.year < max_ - test_window]
    test = data.loc[data.year >= max_ - test_window]
    return train, test


if name == 'main':
    population_data = pd.read_excel(path_join(DATA_DIR, 'Country-Population.xls'))
    melted = melt_data(data=population_data.drop(
        ['Country Code', 'Indicator Name', 'Indicator Code'], axis=1,
    ))
    features = feature_extraction(melted, id_column='Country Name')

    features.to_csv(path_join(DATA_DIR, 'features.csv'), index=False)
```

Requirements.py

```
numpy==1.16.3
pandas==0.24.2
xlrd==1.2.0
lightgbm==2.2.3
scikit-learn==0.21.1
tqdm==4.32.1
```

Timeit.py

```
import time
import sys


def format_time(total_time: float):
    """
    Format imput time to format: '{hours}h {minutes}m {seconds}s'
    :param total_time: time in seconds
    :return: str
    """
    h = total_time // 3600
    min_ = (total_time - h * 3600) // 60
    sec = total_time - h * 3600 - min_ * 60
    txt = '{hours}{minutes}{seconds:.2f}sec'.format(
        hours   = '{:.0f}h '.format(h) if h else '',
        minutes = '{:.0f}min '.format(min_) if min_ else '',
        seconds = sec,
    )
    return txt


class timeit_context:
    """
    Context manager. Print time of block's execution with message: msg
    :param msg: message to print with the time
    :param output: a file-like object (stream); defaults to the current sys.stdout.
    :return:
    """

    def init(self, msg: str = None, output=sys.stdout):
        if msg:
            self.message = msg
        else:
            self.message = 'Time to execute block'
        self.output = output
        self.start_time = time.time()

    def enter(self):
        return self

    def exit(self, exc_type, exc_val, exc_tb):
        if self.output:
            print('{msg}:\t{time}'.format(
```

```
            msg=self.message, time=format_time(time.time() - self.start_time)), file=self.output,
        )
```

Train_test.py

```
import lightgbm
import pickle
import pandas as pd
import numpy as np

from os.path import join as path_join
from sklearn.ensemble import RandomForestRegressor
from typing import Any

from config import DATA_DIR, METRIC_DIR, MODELS_DIR
from metric import rmsle
from timeit import timeit_context


def validation(data, model: Any, min_train_window: int = 20, verbose=False, output_file: str =
'output.txt') -> np.array:
    """ Train model """
    mean_error = []
    for year in range(data.year.min() + min_train_window, data.year.max() + 1, 1):
        train = data[data['year'] < year].drop(labels=['Country Name', 'year'], axis=1)
        val = data[data['year'] == year].drop(labels=['Country Name', 'year'], axis=1)

        x_train, x_val = train.drop(['population'], axis=1), val.drop(['population'], axis=1)
        y_train, y_val = train['population'].values, val['population'].values

        model.fit(x_train, np.log1p(y_train))
        preds = np.expm1(model.predict(x_val))
        mean_error.append(rmsle(ytrue=y_val, ypred=preds))
        if verbose:
            output = f'Year {year} - Error {mean_error[-1]:.5f}'
            print(output)
            with open(output_file, 'a') as file:
                file.write(output + '\n')
    if verbose:
        output = f'Mean Error = {np.mean(mean_error):.5f}'
        print(output)
        with open(output_file, 'a') as file:
            file.write(output + '\n')
    return np.array(mean_error)
```

```
def baseline(data, min_train_window: int = 20, verbose=False, output_file: str = 'output.txt') ->
np.array:
    """ Train model """
    mean_error = []
    for year in range(data.year.min() + min_train_window, data.year.max() + 1, 1):
        val = data[data['year'] == year]

        mean_error.append(rmsle(ytrue=val['population'].values, ypred=val['last_year_pop'].values))
        if verbose:
            output = f'Year {year} - Error {mean_error[-1]:.5f}'
            print(output)
            with open(output_file, 'a') as file:
                file.write(output + '\n')

    if verbose:
        output = f'Mean Error = {np.mean(mean_error):.5f}'
        print(output)
        with open(output_file, 'a') as file:
            file.write(output + '\n')
    return np.array(mean_error)


if name == 'main':
    features = pd.read_csv(path_join(DATA_DIR, 'features.csv'))

    # model = RandomForestRegressor(n_estimators=1000, n_jobs=-1, random_state=5322)
    model = lightgbm.LGBMRegressor(n_estimators=1000, learning_rate=0.01, seed=41)
    with timeit_context("Validation time"):
        metrics = validation(
            data=features,
            model=model,
            min_train_window=20,  # years
            verbose=True,
            output_file=path_join(METRIC_DIR, 'LGB_full_features.txt'),
        )
    pickle.dump(model, open(path_join(MODELS_DIR, 'lgb.pkl'), 'wb'))
```